A local approach to estimation in discrete loglinear models

By Hélène Massam*and Nanwei Wang
Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada

April 22, 2015

Abstract

We consider two connected aspects of maximum likelihood estimation of the parameter for high-dimensional discrete graphical models: the existence of the maximum likelihood estimate (mle) and its computation.

When the data is sparse, there are many zeros in the contingency table and the maximum likelihood estimate of the parameter may not exist. Fienberg and Rinaldo (2012) have shown that the mle does not exists iff the data vector belongs to a face of the so-called marginal cone spanned by the rows of the design matrix of the model. Identifying these faces in high-dimension is challenging. In this paper, we take a local approach : we show that one such face, albeit possibly not the smallest one, can be identified by looking at a collection of marginal graphical models generated by induced subgraphs $G_i, i = 1, \ldots, k$ of G. This is our first contribution.

Our second contribution concerns the composite maximum likelihood estimate. When the dimension of the problem is large, estimating the parameters of a given graphical model through maximum likelihood is challenging, if not impossible. The traditional approach to this problem has been local with the use of composite likelihood based on local conditional likelihoods. A more recent development is to have the components of the composite likelihood be marginal likelihoods centred around each v. We first show that the estimates obtained by consensus through local conditional and marginal likelihoods are identical. We then study the asymptotic properties of the composite maximum likelihood estimate when both the dimension of the model and the sample size N go to infinity.

*H. Massam gratefully acknowledges support from NSERC Discovery Grant No A8947.

Key words: Existence of the maximum likelihood estimate, discrete graphical models, distributed maximum likelihood, faces of the feasible polytope, "large p, large$N$" asymptotics. *AMS 2000 Subject classifications.* 62H17 (Primary), 62M40.

1 Introduction

Let $V = \{1, \ldots, p\}$ be a finite index set. We consider N individuals that we classify according to criteria or variables $X_v, v \in V$. We assume that for each $v \in V$, X_v take its values in a finite set I_v. Let $G = (V, E)$ be an undirected graph where E is the set of undirected edges $(i, j) \in V \times V$. We assume that independences and conditional independences between the variables $X_v, v \in V$ are represented by G in the following way:

$$X_{v_1} \perp X_{v_2} \mid X_{V \setminus \{v_1, v_2\}} \text{ if } (v_1, v_2) \notin E.$$

Thus the distribution of $X = (X_v, v \in V)$ belongs to a discrete graphical model Markov with respect to G. The data are gathered in a p-dimensional contingency table with cells $I = \prod_{v \in V} I_v$. The parameters of this model are the cell probabilities or, equivalently, the loglinear parameters θ in (1.1) below if we write the density of the cell counts under its natural exponential family form

$$f(t; \theta) = \exp\{\langle \theta, t \rangle - Nk(\theta)\} \, . \tag{1.1}$$

Here t is a vector of marginal cell counts and $\langle \theta, t \rangle$ denotes the inner product of $t = t(x)$ and the canonical loglinear parameter θ. Discrete loglinear models are widely used in many scientific areas and two aspects of these models have been the topic of much research. The first is the existence of the maximum likelihood estimate (henceforth abbreviated mle) of the parameters. And the second is that of the computation or approximate computation of the mle. These two aspects are connected as we shall see below. Our contribution in this paper is two-fold and concerns these two aspects.

The mle of the parameter is said to exist if all cell probability estimates are strictly positive. The reader is referred to Fienberg & Rinaldo (2007) and Fienberg & Rinaldo (2012) for a complete list of references on the topic and a most interesting historical account of the developments. It has been known for a long time (see Birch, 1963 and Haberman, 1974) that the nonexistence of the mle is due to the presence of zeros in the contingency table. Zeros can exist even when p is small, but they occur particularly often when p is large and the sample size N is relatively small. Whether the mle exists or not is determined by the position of the data vector t in the convex hull of the support of the measure μ generating the hierarchical loglinear model (1.1). Eriksson et al. (2006)

were the first to express a necessary and sufficient condition for the existence of the mle in these terms. They showed that the mle exists if and only if the data vector belongs to the interior of what they call the marginal cone, i.e., the cone C generated by the convex hull of the support of μ. Fienberg and Rinaldo (2012) showed that this necessary and sufficient condition is valid for all sampling schemes (Poisson, multinomial or product multinomial). The marginal cone C is a polyhedral cone and, thus, in order to determine whether the mle exists, one has to identify whether the data vector belongs to one of its faces. In the supplementary material of their paper, Fienberg and Rinaldo (2012) gave several linear programming algorithms to identify the smallest face of the cone containing the data vector. Practically, however, these algorithms cannot be implemented in high dimensions.

Our first contribution in this paper is to provide, for high-dimensional problems, a way to identify a face of C containing the data vector t. To do so, we take a local approach. We consider a finite collection of subgraphs G_{A_i} of G induced by subsets A_i, $i = 1, \ldots, k$ of V. We solve each local problem, that is, we find the smallest face F_{A_i} containing the vector of marginal cell counts t_{A_i} in the marginal model Markov with respect to G_{A_i}. This face can be extended to a face F_i of C containing t. In Theorem 3.1, we show that $\cap_{i=1}^k F_i$ is a face of the marginal cone C of the global model containing the data vector t. The face $\cap_{i=1}^k F_i$ is not necessarily the smallest one containing t, but knowing that t belongs to $\cap_{i=1}^k F_i$ tells us that the mle does not exist and gives us a model of dimension smaller than the dimension of $\mathcal{M}$ for which we can attempt to evaluate the mle. Of course, this is true only if $\cap_{i=1}^k F_i$ is not equal to the relative interior of C. If it is, our procedure yields no information. In our simulations, if the data belonged to a face of C, we never had this situation. In fact, we found that the face containing t was always equal to the intersection $\cap_{i=1}^k F_i$. This is probably due to the fact that a simulated data vector will rarely fall on a face of small dimension which would not show up as $\cap_{i=1}^k F_i$.

Our second contribution in this paper concerns the composite maximum likelihood estimates of θ in (1.1). The mle of θ is that value of the parameter θ that maximizes the likelihood or, equivalently, the loglikelihood function $l(\theta) = \langle \theta, t \rangle - Nk(\theta)$. The log partition function or cumulant generating function $k(\theta)$ is usually intractable when the dimension of the model is large. As a consequence, even though the likelihood function is a convex function of θ, the traditional convex optimization methods using the derivative of the likelihood cannot be used. Approximate techniques such as variational methods (see Jordan et al., 1999, Wainwright and Jordan, 2008) or MCMC techniques (see Geyer, 1991, Snijders 2002) have been developed in recent years. More recently still, work has been done on a third type of approximate techniques based on the maximization of composite likelihoods (Lindsay, 1988). For a given graphical model and a given data set $x^{(1)}, \ldots, x^{(N)}$, a composite likelihood is typically the product of local conditional likelihoods, coming from

the local conditional probability of X_v given $X_{\mathcal{N}_v}$, which we can write as

$$L^{PS}(\theta) = \prod_{v \in V} \prod_{k=1}^{N} p(X_v = x_v^{(k)} | X_{\mathcal{N}_v} = x_{\mathcal{N}_v}^{(k)}; \theta) \tag{1.2}$$

where $\mathcal{N}_v$ indices the set of neighbours of v in G. For work using this type of techniques applied to discrete graphical models, the reader may refer to Liu and Ihler (2012) and references therein.

In the last two years, for Gaussian graphical models, Wiesel and Hero (2012) and Meng et al. (2013, 2014) have introduced composite likelihood estimation where the composite likelihood is the product of local convex "relaxed" marginal (rather than conditional) likelihoods coming from $p(X_v, X_{\mathcal{N}_v})$ rather than from $p(X_v \mid X_{\mathcal{N}_v})$. For discrete graphical models, Mizrahi et al. (2014a) have proposed a similar marginal approach, taking a clique of G and its neighbourhood. In the papers mentioned above using the composite likelihood, from either marginal or conditional local likelihoods, the global composite mle is obtained by "consensus", i.e. by combining the local mle's obtained from the various v-local likelihoods. The value of the global composite mle is therefore a function of the values of the mle of the v-local likelihoods.

In this paper, we extend the marginal approach of Meng et al. (2013, 2014) to discrete graphical models and give two results. The first, Theorem 4.1 states that the composite mle obtained through this marginal approach is equal to the composite mle obtained through the conditional approach. We conclude that there is therefore no point in using the more computationally complex marginal approach to compute the composite mle. Our second result, Theorem 5.1 gives the rate of convergence of the composite mle to the true value θ^* of the parameter when both p and N go to infinity. In Theorem 5.2, we give the rate of convergence of the global mle, under the same conditions and compare the two rates of convergence.

The existence and computation of the mle are of course connected, if a local mle does not exist, the composite mle cannot be computed by consensus and probably does not exist. An optimization routine may not signal that the mle does not exist and will usually give a fallacious number for the local mle In this case, running a convex optimization routine on each local problem without looking at the details of each maximization may lead to an inaccurate estimate (and erroneously lead to think that the two composite mle, obtained from local marginal and conditional likelihoods, are different).

The remainder of this paper is organized as follows. In the next section, we introduce our notation for the discrete graphical model. In Section 3, we prove our result on the identification of a face containing the data vector. In Section 4, we study the local marginal

likelihood estimator and its relationship to the local conditional likelihood estimator. In Section 5, we give its asymptotic properties. The proof of the theorems are given in the Appendix.

2 Preliminaries

2.1 Discrete graphical and hierarchical loglinear models

Let p, V and $X = (X_v,\ v \in V)$ be as described in Section 1 above. If N individuals are classified according to the p criteria, the resulting counts are gathered in a contingency table such that

$$I = \prod_{v \in V} I_v$$

is the set of cells $i = (i_v,\ v \in V)$. For $D \subset V$, i_D denotes the marginal cell $i_D = (i_v, v \in D)$ with $i_v \in I_v$. Let $\mathcal{D}$ be a family of non empty subsets of V such that $D \in \mathcal{D}$, $D_1 \subset D$ and $D_1 \neq \emptyset$ implies $D_1 \in \mathcal{D}$. In order to avoid trivialities we assume $\cup_{D \in \mathcal{D}} D = V$. The family $\mathcal{D}$ is called the generating class of the hierarchical loglinear model. We denote by $\Omega_{\mathcal{D}}$ the linear subspace of $y \in \mathbb{R}^I$ such that there exist functions $\theta_D \in \mathbb{R}^I$ for $D \in \mathcal{D}$ depending only on i_D and such that $y = \sum_{D \in \mathcal{D}} \theta_D$, that is

$$\Omega_{\mathcal{D}} = \{y \in \mathbb{R}^I :\ \exists \theta_D \in \mathbb{R}^I, D \in \mathcal{D} \text{ such that } \theta_D(i) = \theta_D(i_D) \text{ and } y = \sum_{D \in \mathcal{D}} \theta_D\}$$

The hierarchical model generated by $\mathcal{D}$ is the set of probabilities $p = (p(i))_{i \in I}$ on I such that $p(i) > 0$ for all i and such that $\log p \in \Omega_{\mathcal{D}}$.

The class of discrete graphical models Markov with respect to an undirected graph G is a subclass of the class of hierarchical discrete loglinear models. Indeed, let $G = (V, E)$ be an undirected graph where V is the set of vertices and $E \subset V \times V$ denotes the set of undirected edges. We say that the distribution of X is Markov with respect to G if $(v_1, v_2) \notin E$ implies

$$X_{v_1} \perp X_{v_2} |\ X_{V \setminus \{v_1, v_2\}}.$$

Let $\mathcal{D}$ be the set of all cliques (not necessarily maximal) of the graph G. If the distribution of $X = (X_1, \ldots, X_p)$ is Multinomial$(1, p(i), i \in I)$ Markov with respect to the graph G, and if we assume that all $p(i), i \in I$, are positive, then, by the Hammersley-Clifford theorem, $\log p(i)$ is a linear function of parameters dependent on the marginal cells $i_D, D \in \mathcal{D}$ only, and therefore the graphical model is a hierarchical loglinear model with generating set the set $\mathcal{D}$ of cliques of G. The reader is referred to Darroch & Speed (1983), Lauritzen

(1996) or Letac & Massam (2012) for a detailed description of the hierarchical loglinear model and the subclass of discrete graphical loglinear models.

We now set our notation and recall some basic results for discrete hierarchical loglinear models. The following notation and results can be found in Letac & Massam (2012) and the corresponding supplementary file.

Among all the values that X_v can take in $I_v, v \in V$, we call one of them 0. For a cell $i \in I$, we define its support $S(i)$ as

$$S(i) = \{v \in V \ ; \ i_v \neq 0\}$$

and we define also the following subset J of I

$$J = \{j \in I, \quad S(j) \in \mathcal{D}\}. \tag{2.1}$$

From here on, we will call this set the J-set of the model. For $i \in I$ and $j \in J$, we define the symbol

$$j \lhd i$$

to mean that $S(j)$ is contained in $S(i)$ and that $j_{S(j)} = i_{S(j)}$. The relation $\lhd$ has the property that if $j, j' \in J$ and $i \in I$, then

$$j \lhd j' \ \text{ and } \ j' \lhd i \Rightarrow j \lhd i.$$

The log linear parametrization that we use for the multinomial is the so-called baseline parametrization with general expression, for $i \in I, S(i) = E \subset V$,

$$\theta_i = \sum_{F \subset E} (-1)^{|E|-|F|} \log p(i_F, 0_{V \setminus F}) \ . \tag{2.2}$$

With the notation above, in Proposition 2.1 of Letac and Massam (2012), it is shown that for $i \notin J, \ \theta_i = 0$ and that

$$\begin{aligned} \theta_j &= \sum_{j' \in J, \ j' \lhd j} (-1)^{|S(j)|-|S(j')|} \log \frac{p(j')}{p(0)}, \ j \in J \\ \log p(i) &= \theta_0 + \sum_{j \in J, j \lhd i} \theta_j, \ i \in I \qquad (2.3) \\ \log p(0) &= \theta_0. \qquad (2.4) \end{aligned}$$

One then readily derives the density of the multinomial $\mathrm{M}(N, p(i), i \in I)$ of the cell counts $\underline{n} = (n(i), i \in I)$, Markov with respect to G to be, up to a multiplicative constant, equal to

$$f(t; \ \theta) = \exp\{\langle t, \theta \rangle - Nk(\theta)\}, \ \theta \in R^J \tag{2.5}$$

with $\theta = (\theta_j, j \in J),\ t = t(\underline{n}) = (t(j),\ j \in J)$ where $t(j) = n(j_{S(j)})$ are the $j_{S(j)}$-marginal cell counts and

$$k(\theta) = \log\Big(\sum_{i\in I} \exp \sum_{j\in J, j\lhd i} \theta_j\Big) = \log\Big(1 + \sum_{i\in I\setminus\{0\}} \exp \sum_{j\in J, j\lhd i} \theta_j\Big)\ . \tag{2.6}$$

For $\theta \in R^J$, these distributions form a natural exponential family of dimension J generated by a measure μ which we will now identify. Let $e_j, j \in J$ be the canonical basis of R^J and, for $i \in I$, let

$$f_i = \sum_{j\in J, j\lhd i} e_j.$$

Then (2.3) and (2.4) can be written in matrix form as

$$\log p \quad = \quad A\tilde{\theta} \tag{2.7}$$

where $\tilde{\theta}^t = (\theta_0, \theta^t)$, A is an $(|I|) \times (1 + |J|)$ matrix. We call A the design matrix of the model. The rows of A are indexed by $i \in I$ and equal to $\tilde{f}_i^{\,t} = (1, f_i^t) \in \mathbb{R}^{J+1}$. It is immediate to see that the Laplace transform of the generating measure μ is

$$e^{k(\theta)} = \sum_{i\in I} e^{\langle\theta, f_i\rangle}$$

and therefore the measure μ generating (2.5) is

$$\mu(dx) = \sum_{i\in I} \delta_{f_i}(x). \tag{2.8}$$

This exponential family is concentrated on the convex hull of $f_i, i \in I$, which is a bounded set of $\mathbb{R}^J$, and therefore the set of parameters θ for which L is finite is the whole space $\mathbb{R}^J$. From the definition of X, $f_i, i \in I$ and $t = (t(j_{S(j)}, j \in j)$, it is easy to see that $(N, t(j),\ j \in J)^t = A^t n = \sum_{i\in I} n(i)\tilde{f}_i$ and the vector of sufficient statistics t, which we also write as $t = t_J$ to emphasize its length, is such that

$$\frac{t_J}{N} = \Big(\frac{t(j)}{N}, j \in J\Big)^t = \sum_{i\in I\setminus\{0\}} \frac{n(i)}{N} f_i = \sum_{i\in I} \frac{n(i)}{N} f_i \tag{2.9}$$

belongs to the convex hull of $(f_i)_{i\in I}$. The $(f_i)'s$ are the extreme points of the closure of the convex hull of the $f_i, i \in I$.

Fienberg and Rinaldo (2012), Theorem 3, show that the mle of θ in (2.5) exists and is unique if and only if the canonical statistic vector t belongs to the relative interior of the

cone C with apex f_0 and generated by $f_i, i \in I \setminus \{0\}$. Therefore, if the mle does not exist, the data vector t must belong to one of the facets of C. Following Eriksson et al. (2006) and Fienberg and Rinaldo (2012), we will call C the marginal cone of the model. The reader is referred to Letac and Massam (2012) for examples of the notions given above.

3 Faces containing the data vector t

3.1 Finding the smallest face containing t

Let C be a closed convex polyhedral cone. A set $F \subset C$ is said to be a face of C if for all $x \in F$, any decomposition of the form $x = y + z$ with $y, z \in C$ implies $y, z \in F$. Given $g \in R^J$, the inequality $\langle g, x\rangle \geq 0$ is said to be valid for C if it holds for every $x \in C$. Then the set

$$F_g = \{x \in C : \ \langle g, x\rangle = 0\}$$

is called the face of C governed by g. Every face of C arises in this manner. There is only one face of dimension 0 and that is $\{f_0\}$. The faces of dimension 1 are the extreme rays $\{\lambda f_i, \ \lambda > 0\}$ for each $f_i, i \in I \setminus \{0\}$. Since a convex set is the convex hull of its extreme points, a face F of C will be defined by a set

$$f_i, i \in \mathcal{F} \subset I.$$

To identify the smallest face F of C containing the data vector $t \in R^J$, we will have to identify the subset $\mathcal{F}$ of I defining that face or equivalently we will have to identify a $g \in R^J$ such that

$$\begin{aligned} \langle g, f_i\rangle &= 0, \ \forall i \in \mathcal{F} \\ \langle g, f_i\rangle &> 0, \ \forall i \in \mathcal{I} \setminus \mathcal{F} \end{aligned}$$

Let I_+ be the subset of $\mathcal{I}$ with strictly positive cell counts $n(i) > 0$. We have the following lemma which shows that $\mathcal{F} \supset I_+$.

Lemma 3.1 *The sufficient statistic t belongs to the face F_g of C governed by g if and only if $f_i \in F_g$ for all $i \in I_+$.*

The proof is obvious if we write that $t \in F_g \Leftrightarrow \langle t, g\rangle = 0 \Leftrightarrow \sum_{i\in\mathcal{I}_+} \frac{n(i)}{N}\langle f_i, g\rangle = 0 \Leftrightarrow \langle f_i, g\rangle = 0 \ \forall i \in I_+$. The face F_g may contain additional f_i's. Recall that the design matrix A given in (2.7) has rows $(1, f_i^t)$. Let B be the $|I| \times |J|$ matrix with rows $f_i^t, i \in I$. Let B_+ be the matrix obtained from B by keeping the rows indexed by I_+ only, and let B_0

be the matrix obtained by keeping the rows in $I \setminus I_+$ only. Fienberg and Rinaldo (2012) showed that $\mathcal{F} \setminus I_+$ could be identified by solving the linear program

$$\begin{aligned} &\text{Max } ||Bg||_0 \\ &\text{s.t. } B_+ g = 0 \\ &B_0 g \geq 0, \end{aligned} \tag{3.1}$$

where $||.||_0$ is the zero norm in R^J. This is a non-convex program which is, however, easily solved by a sequence of linear programming relaxations (see programs (4) and (6) in Fienberg and Rinaldo (2012), supplementary material). We adopted their approach to find the smallest face containing the data vector t whenever the dimension is small enough for the program to run. We found that we could use this program only for models with up to 16 vertices.

3.2 Splitting the global model into smaller models

Since it is difficult or impossible to implement the program (3.1) in high-dimension, we now consider a collection of smaller models and we show that the combination of these smaller models yields information on the global model.

Suppose that X follows a model M Markov with respect to the undirected graph $G = (V, E)$. Let $A \subset V$ and let G_A be the graph induced by A. Let M_A be the model Markov with respect to G_A; this is also a multinomial model. The generating set $\mathcal{D}_A$ of M_A is a subset of $\mathcal{D}$. Let

$$J_A = \{j \in J, \quad S(j) \in \mathcal{D}_A\}. \tag{3.2}$$

Then $t_{J_A} = (t(j) = n(j_{S(j)}),\ j \in J_A)$ is the canonical data vector of M_A. Let C_A be the marginal cone of M_A. Clearly C_A is generated by $f_{i_A,J_A},\ i_A \in I_A$ where $f^t_{i_A,J_A} = (f_{ij}, j \in J_A)$ is made up of the components of $f_i, i \in I$ that are indexed by J_A. We may have f_{i_A,J_A} coming from different f_i's but clearly we keep only one copy. The following lemma shows how to extend a face of C_A into a face of C.

Lemma 3.2 *For $g_A \in R^{J_A}$, let $F_{g_A} = \{x \in C_A :\ \langle g_A, x\rangle = 0\}$ be a face of C_A containing t_{J_A}. Let $g_1 = (g_A, 0 \ldots, 0) \in R^J$ be the vector of R^J obtained from g_A by setting the remaining variables to 0. Then*

$$F_{g_1} = \{x \in C :\ \langle g_1, x\rangle = 0\}$$

is a face of C containing t.

Proof. Let $\mathcal{F}_A$ denote the index set of the f_{i_A} defining F_{g_A}. We have

$$\begin{aligned} \langle g_A, f_{i_A,J_A} \rangle &= 0,\ i_A \in \mathcal{F}_A \\ \langle g_A, f_{i_A,J_A} \rangle &> 0,\ i_A \notin \mathcal{F}_A \\ \langle g_A, t_{J_A} \rangle &= 0. \end{aligned}$$

Writing $i \in I$ as $i = (i_A, i_{A^c})$ and $f_i = (f_{i,J_A}, f_{i,J_{A^c}})$, we have

$$\begin{aligned} \langle g_1, f_i \rangle &= \langle g_A, f_{i,J_A} \rangle + \langle 0, f_{i,J_{A^c}} \rangle = 0,\ i \text{ such that } i_A \in \mathcal{F}_A \\ \langle g_1, f_i \rangle &= \langle g_A, f_{i,J_A} \rangle + \langle 0, f_{i,J_{A^c}} \rangle > 0,\ i \text{ such that } i \notin \mathcal{F}_A \\ \langle g_1, t \rangle &= \langle g_A, t_{J_A} \rangle + \langle 0, t_{J_{A^c}} \rangle = 0. \end{aligned}$$

It follows immediately that F_{g_1} is a face of C containing t and it is defined by $f_i,\ i_A \in \mathcal{F}_A$. □

We now use this lemma for a finite number of smaller models and in doing so, obtain an even smaller face containing the data vector t. Let M be the model Markov with respect to $G = (V, E)$. Let $G_{A_l},\ l = 1 \ldots, k$ be a collection of subgraphs induced by subsets $A_l,\ l = 1, \ldots, k$ of V. Let $\mathcal{D}_{A_l}$ be the generating set for the model M_{A_l} Markov with respect to G_{A_l} and let t_{A_l} be its canonical statistic.

Theorem 3.1 *Let $A_l,\ l = 1, \ldots, k$ be a collection of subsets of V. Let g_{A_l} define faces of C_{A_l} containing $t_{J_{A_l}},\ l = 1, \ldots, k$ respectively, let g_l be the vectors of R^J obtained by completing g_{A_l} with zeros and let $F_l = F_{g_l}$ be the corresponding faces. Then*

$$g = \sum_{l=1}^{k} g_l$$

defines a face $\cap_{l=1}^{k} F_l$ of C containing t.

Proof. We give the proof for $k = 2$. If $f_i \in F_1 \cap F_2$, we have that $\langle f_i, g_1 \rangle = \langle f_i, g_2 \rangle = 0$ and therefore $\langle f_i, g \rangle = 0$. Moreover, if $f_i \notin F_1 \cap F_2$, either $\langle f_i, g_1 \rangle > 0$ or $\langle f_i, g_2 \rangle > 0$, and therefore $\langle f_i, g \rangle > 0$. It follows that $F_g = \{x \in C : \langle g, x \rangle = 0\}$ is a face of C. Moreover, $\langle g, t \rangle = \langle g_1, t \rangle + \langle g_2, t \rangle = 0$ and therefore $t \in F_g$ which clearly is equal to $F_1 \cap F_2$. □

Remark 3.1 *In practice, when applying the theorem above, we choose the collection A_l so that $\mathcal{D} = \cup_{l=1}^{k} \mathcal{D}_{A_l}$ with $A_l, l = 1, \ldots, k$ as large as possible but small enough that we can compute the smallest face containing $t_{J_{A_l}}$. In experiments similar to that presented in Section 3.4 below, we nearly always found that $\cap_{l=1}^{k} F_l$ was the smallest face containing t.*

3.3 A numerical experiment

We illustrate our results with a numerical experiment. We consider the graphical model with G equal to the four-neighbour 4×4 lattice with binary data so that $|I| = 2^{16}$ and $|J| = 40$. We generated the data vector

$$t = [1\ 2\ 7\ 5\ 4\ 5\ 2\ 7\ 7\ 3\ 2\ 10\ 2\ 8\ 1\ 5\ 0\ 0\ 1\ 1\ 4\ 1\ 4\ 1\ 3\ 1\ 2\ 2\ 0\ 7\ 3\ 1\ 1\ 2\ 2\ 1\ 5\ 2\ 0\ 0].$$

We let $A_i, i = 1, 2, 3$ be the subsets of V comprising the vertices in the i-th and $(i+1)$-th row of the lattice. We find that the face F_{A_1} of C_{A_1} containing $t_{J_{A_1}}$ is of dimension 15 and the expanded face F_1 is therefore of dimension $15 + 22 = 37$. The corresponding g_1 is

$$g_1 = 10\ \Big(0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ -1\Big).$$

The face F_{A_2} is of dimension 13 while F_2 is of dimension $13 + 22 = 35$. The corresponding vector is

$$g_2 = 10\ \Big(0,\ 0,\ 1,\ 1,\ 0,\ 1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ -1,\ 1,\ -1,\ -1,\ 0,\ -1\Big)$$

The face F_{A_3} is of dimension 11 and F_3 is of dimension $11 + 22 = 33$ with

$$g_3 = 10\ \Big(0,\ 1,\ 1,\ 0,\ 1,\ 0,\ 1,\ 1,\ -1,\ 0,\ 0,\ 0,\ -1,\ -1,\ -1,\ -1,\ 1,\ 1\Big).$$

In this particular case, after computing the facial set of F, the smallest face of the global model containing t, we find that actually $F = F_1 \cap F_2 \cap F_3$ which is of dimension 28.

4 The conditional and marginal composite maximum likelihood estimators

When the dimension of the discrete graphical model is large, computing the maximum likelihood estimate of θ in (2.5) is challenging, if not impossible. As mentioned in the introduction, a recent approach to this problem has been local with the use of a composite likelihood which is equal to the product, over all vertices $v \in V$, of the local conditional likelihood for X_v given $X_{\mathcal{N}_v}$ where $\mathcal{N}_v$ denotes the set of neighbours of v in G. Recently, for Gaussian high-dimensional graphical models, Wiesel & Hero (2012) and Meng & al. (2013, 2014) worked with a different composite likelihood which is the product, over all vertices $v \in V$, of local marginal likelihoods. In this section, we will first recall the definition of the conditional composite likelihood estimate, then extend the marginal

composite likelihood to discrete graphical models and finally show that the maximum likelihood estimates obtained from these two types, conditional and marginal, of local models are in fact identical and thus the composite likelihood obtained by consensus from these two types of likelihood are equal. Since the computational complexity of the marginal computations is exponential in the number of vertices in the neighbourhood of v while the conditional computations are linear in this number, there is no advantage in working with marginal composite likelihoods.

4.1 The conditional composite likelihood function

We first define the standard conditional composite likelihood function. For $i = (i_v, v \in V)$, let $i^{(1)}, \ldots, i^{(N)}$ be a sample of size N from the distribution of X Markov with respect to G. We recall that the global likelihood function is

$$L(\theta) \propto \prod_{k=1}^{N} p(X_v = i_v^{(k)}, v \in V) = \exp\{\langle \theta, t \rangle - Nk(\theta)\} \tag{4.1}$$

where $k(\theta)$ is as in (2.6).

For a given vertex $v \in V$, let $\mathcal{N}_v$ the set of neighbours of v in the given graph G. The composite likelihood function based on the local conditional distribution of X_v given $X_{V\setminus\{v\}}$ or equivalently, due to the Markov property, the conditional distribution of X_v given its neighbours $X_{\mathcal{N}_v}$ is $L^{PS}(\theta) = \prod_{v\in V} L^{PS_v}(\theta)$ where

$$L^{PS_v}(\theta) = \prod_{k=1}^{N} p(X_v = i_v^{(k)} | X_{\mathcal{N}_v} = i_{\mathcal{N}_v}^{(k)}; \theta) \tag{4.2}$$

and the superscript PS stands for "pseudo-likelihood", the name often given to the conditional composite lilelihood (Besag, 1974). As given by (2.3), for a given cell i, we have

$$\begin{aligned} \log p(i) &= \log p(X_v = i_v, v \in V) = \theta_0 + \sum_{j \lhd i} \theta_j \\ &= \theta_0 + \sum_{j \lhd i,\ S(j) \subseteq v \cup \mathcal{N}_v, S(j) \not\subseteq \mathcal{N}_v} \theta_j + \sum_{j \lhd i,\ S(j) \subseteq \mathcal{N}_v} \theta_j + \sum_{j \lhd i,\ S(j) \not\subseteq v \cup \mathcal{N}_v} \theta_j \end{aligned}$$

The set J is as defined in (2.1) for the global model. Let

$$J^{PS_v} = \{j \in J \mid S(j) \subseteq v \cup \mathcal{N}_v, S(j) \not\subseteq \mathcal{N}_v\} = \{j \in J \mid v \in S(j)\}.$$

Then for $i_v \neq 0$, we have

$$p(X_v = i_v |\ X_{\mathcal{N}_v} = i_{\mathcal{N}_v}) = p(X_v = i_v |\ X_{V\setminus\{v\}} = i_{V\setminus\{v\}}) = \frac{p(X_V = i_V)}{p(X_{V\setminus\{v\}} = i_{V\setminus\{v\}})}$$

$$= \frac{e^{\theta_0 + \sum_{j \lhd i,\ j \in J^{PS_v}} \theta_j + \sum_{j \lhd i,\ S(j) \subseteq \mathcal{N}_v} \theta_j + \sum_{j \lhd i,\ S(j) \not\subseteq v \cup \mathcal{N}_v} \theta_j}}{\sum_{k \in I |\ k_{V\setminus\{v\}} = i_{V\setminus\{v\}}} \left(e^{\theta_0 + \sum_{j \lhd k,\ j \in J^{PS_v}} \theta_j + \sum_{j \lhd k,\ S(j) \subseteq \mathcal{N}_v} \theta_j + \sum_{j \lhd k,\ S(j) \not\subseteq v \cup \mathcal{N}_v} \theta_j} \right)}$$

$$= \frac{e^{\sum_{j \lhd i,\ j \in J^{PS_v}} \theta_j}}{1 + \sum_{k \in I |\ k_{V\setminus\{v\}} = i_{V\setminus\{v\}},\ k_v \neq 0} e^{\sum_{j \lhd k,\ j \in J^{PS_v}} \theta_j}} \tag{4.3}$$

and

$$p(X_v = 0 |\ X_{V\setminus\{v\}} \ = \ i_{V\setminus\{v\}}) = \frac{1}{1 + \sum_{k \in I |\ k_{V\setminus\{v\}} = i_{V\setminus\{v\}},\ k_v \neq 0} e^{\sum_{j \lhd k,\ j \in J^{PS_v}} \theta_j}} \tag{4.4}$$

Equality (4.3) is due to the fact that the set of $j \in J$ such that $j \lhd k,\ S(j) \not\subseteq v \cup \mathcal{N}_v$, is the same whether $k_v = i_v$ or $k_v \neq i_v$ and therefore the term $e^{\theta_0 + \sum_{j \lhd k,\ S(j) \not\subseteq k_v \cup \mathcal{N}_v} \theta_j}$ cancels out at the numerator and the denominator. The same goes for the set of $j \in J$ such that $j \lhd k,\ S(j) \subseteq \mathcal{N}_v$.

Remark 4.1 *In the equation above, we worked with $p(X_v|X_{V\setminus\{v\}})$ rather than with $P(X_v|X_{\mathcal{N}_v})$, though the two are equal, in order to emphasize that the parameter*

$$\theta^{PS_v} = (\theta_j,\ j \in J^{PS_v}),\ \ v \in V \tag{4.5}$$

of the v-th component L^{PS_v} of conditional composite distribution is a subvector of θ, the parameter of the global likelihood function.

We now define the two-hop conditional composite likelihood function.

Definition 4.1 *For a given $v \in V$, we will say that $\mathcal{M}_v$ is a one-hop neighbourhood of v if it comprises v and its immediate neighbours in G, i.e. if $\mathcal{M}_v = \{v\} \cup \mathcal{N}_v$. We will say that $\mathcal{M}_v$ is a two-hop neighbourhood if it comprises v, its immediate neighbours and the neighbours of the immediate neighbours in G. We use the notation*

$$\mathcal{N}_{2v} = \mathcal{M}_v \setminus \Big(\{v\} \cup \mathcal{N}_v \Big)$$

to denote the set of neighbours of the neighbours of v. For simplicity of notation, we will denote both the one-hop and two-hop neighbourhoods by $\mathcal{M}_v$.

The two-hop conditional composite likelihood function is $L^{PS_2}(\theta) = \prod_{v\in V} L^{PS_{2,v}}(\theta)$

$$L^{PS_{2,v}}(\theta) = \prod_{k=1}^{N} p(X_v = i_v^{(k)}, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}^{(k)} | X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}^{(k)}). \tag{4.6}$$

The expression of $p(X_v = i_v^{(k)}, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}^{(k)} | X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}^{(k)})$ is the same as (4.3) and (4.4) but with J^{PS_v} replaced by $J^{PS_{2,v}}$ where

$$J^{PS_{2,v}} = \{j \in J \mid S(j) \subseteq \mathcal{M}_v, S(j) \not\subseteq \mathcal{N}_{2v}\}.$$

In a parallel way to Remark 4.1, we note that

$$\theta^{PS_{2,v}} = \{\theta_j, j \in J^{PS_{2,v}}\}$$

is a subvector of $\theta = (\theta_j, j \in J)$, the argument of the global likelihood function.

4.2 The marginal composite likelihood

Let $\mathcal{M}_v$ be the one-hop or two-hop neighbourhood of v. The marginal composite likelihood is the product

$$L^{\mathcal{M}}(\theta) = \prod_{v\in V} \prod_{k=1}^{N} p(X_{\mathcal{M}_v} = i_{\mathcal{M}_v}^{(k)}) = \prod_{v\in V} L^{\mathcal{M}_v}(\theta). \tag{4.7}$$

where $L^{\mathcal{M}_v}(\theta) = \prod_{k=1}^{N} p(X_{\mathcal{M}_v} = i_{\mathcal{M}_v}^{(k)})$. The $\mathcal{M}_v$-marginal model is clearly multinomial and the corresponding data can be read in the $\mathcal{M}_v$-marginal contingency table obtained from the full table. The density of the $\mathcal{M}_v$-marginal multinomial distribution is of the general exponential form

$$f(t^{\mathcal{M}_v};\ \theta^{\mathcal{M}_v}) = \exp\{\langle t^{\mathcal{M}_v}, \theta^{\mathcal{M}_v}\rangle - N k^{\mathcal{M}_v}(\theta^{\mathcal{M}_v})\} \tag{4.8}$$

where $t^{\mathcal{M}_v}$, $\theta^{\mathcal{M}_v}$ and $k^{\mathcal{M}_v}$ are respectively the $\mathcal{M}_v$-marginal canonical statistic, canonical parameter and cumulant generating function.

In order to identify the $\mathcal{M}_v$-marginal model, we first establish the relationship between θ and $\theta^{\mathcal{M}_v}$. In the sequel, the symbol j will be understood to be an element of $I_{\mathcal{M}_v}$ when used in the notation $\theta_j^{\mathcal{M}_v}$ while it will be understood to be the element of J obtained by padding it with entries $j_{V\setminus\mathcal{M}_v} = 0$ when used in the notation θ_j. We now give the general relationship between the parameters of the overall model and those of the $\mathcal{M}_v$-marginal model. Proofs are given in the Appendix.

Lemma 4.1 *Let $\mathcal{M}_v$ be the one-hop or two-hop neighbourhood of $v \in V$. For $j \in J, S(j) \subset \mathcal{M}_v$, the parameter θ_j of the overall model and the parameter $\theta_j^{\mathcal{M}_v}$ of the marginal model are linked by the following:*

$$\theta_j^{\mathcal{M}_v} = \theta_j + \sum_{j' \mid j' \lhd_0 j} (-1)^{|S(j)-S(j')|} \log\left(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v} = j'} \exp \sum_{\substack{k \mid k \lhd i \\ k \not\lhd j'}} \theta_k\right). \tag{4.9}$$

We now want to identify which of the marginal parameters are equal to the corresponding overall parameter and in particular which marginal parameters are equal to 0 when the global parameter is equal to zero. Let $\mathcal{M}_v^c$ denote the complement of $\mathcal{M}_v$ in V. We define the buffer set at v as follows:

$$\mathcal{B}_v = \{w \in \mathcal{M}_v \mid \exists w' \in \mathcal{M}_v^c \text{ with } (w, w') \in E\}. \tag{4.10}$$

We have the following result.

Lemma 4.2 *Let $\mathcal{M}_v$ be the one-hop or two-hop neighbourhood of $v \in V$. For $j \in J, S(j) \subset \mathcal{M}_v$ the following holds:*

(1.) if $S(j) \not\subset \mathcal{B}_v$, then $\theta_j^{\mathcal{M}_v} = \theta_j$,

(2.) if $S(j) \subset \mathcal{B}_v$, then in general $\theta_j^{\mathcal{M}_v} \neq \theta_j$, and (4.9) holds.

Moreover, for $i \in I, S(i) \subset \mathcal{M}_v$,

(3.) If $S(i) \not\subset \mathcal{B}_v$, then $\theta_i^{\mathcal{M}_v} = 0$ whenever $\theta_i = 0$.

From the lemma above, we see that, for $j \in J$ such that $S(j) \subset \mathcal{M}_v, S(j) \not\subset \mathcal{B}_v$, the corresponding global and $\mathcal{M}_v$-marginal loglinear parameters are equal. We see also that for $i \in I$ such that $S(i) \in \mathcal{M}_v, S(i) \not\subset \mathcal{B}_v$, if the loglinear parameter is zero in the global model, it remains zero in the $\mathcal{M}_v$-marginal model.

4.3 A convex relaxation of the local marginal optimization problems

It is clear from (4.9) that even though maximizing the marginal likelihood from (4.8) is convex in $\theta^{\mathcal{M}_v}$, it is not convex in θ. We would therefore like to replace the problem of maximizing (4.8) non convex in θ by a convex relaxation problem. We know from *(1.)* of Lemma 4.2 that $\theta_j^{\mathcal{M}_v} = \theta_j$ for j in the set $\{j \in J : S(j) \subset \mathcal{M}_v,\ S(j) \not\subset \mathcal{B}_v\}$.

We also know from *(3.)* of Lemma 4.2 that if the global model parameter $\theta_i, S(i) \subset \mathcal{M}_v, S(i) \not\subset \mathcal{B}_v$ is equal to zero, then $\theta_i^{\mathcal{M}_v}$ is also equal to zero. Following what has been done for Gaussian graphical models, it is then natural to consider the following graphical model relaxation of the $\mathcal{M}_v$-marginal model.

Let $\mathcal{M}_{l,v}$ index the relaxed hierarchical loglinear model obtained from the $\mathcal{M}_v$-marginal model by keeping interactions given by edges with at least one endpoint in $\mathcal{M}_v \setminus \mathcal{B}_v$ and only those edges, and all interactions in the power set $2^{\mathcal{B}_v}$. The index l takes values $l = 1$ or $l = 2$ when $\mathcal{M}_v$ is respectively the one-hop or two-hop neighbourhood of v. The J-set of this local model is

$$J^{\mathcal{M}_{l,v}} = \{j \in J \mid S(j) \subset \mathcal{M}_v, S(j) \not\subset \mathcal{B}_v\} \cup \{i \in I \mid S(i) \subset \mathcal{B}_v\} \ . \tag{4.11}$$

Let $p^{\mathcal{M}_{l,v}}(X_{\mathcal{M}_v})$ denote the marginal probablity of $X_{\mathcal{M}_v}$ in the $\mathcal{M}_{l,v}$-marginal model. The local estimates of $\theta_j, j \in \{j \in J |\ S(j) \subset \mathcal{M}_v,\ S(j) \not\subset \mathcal{B}_v\}$ are obtained by maximizing the $\mathcal{M}_{l,v}$-marginal loglikelihood

$$L^{\mathcal{M}_{l,v}}(\theta) = \prod_{k=1}^{N} p^{\mathcal{M}_{l,v}}(X_{\mathcal{M}_v} = i_{\mathcal{M}_v}^{(k)}) = \exp\{\langle \theta^{\mathcal{M}_{l,v}}, t^{\mathcal{M}_{l,v}} \rangle - N k^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})\} \tag{4.12}$$

which is a convex maximization problem in

$$\theta^{\mathcal{M}_{l,v}} = (\theta_j, j \in J^{\mathcal{M}_{l,v}}).$$

At this point, we need to make two important remarks.

Remark 4.2 *The vector θ^{PS_v} defined in (4.5) is a subvector of $\theta^{\mathcal{M}_{l,v}}$. Therefore maximizing (4.12) for either $l =$ or $l = 2$ will yield an estimate of θ^{PS_v}.*

Remark 4.3 *The $\mathcal{M}_{l,v},\ l = 1, 2$-marginal model is a hierarchical loglinear model but not necessarily a graphical model. For example, if we consider a four-neighbour lattice and a given vertex v_0 and its four neighbours that we will call $1, 2, 3, 4$ for now, then the generating set of the relaxed $\mathcal{M}_{1,v_0}$-marginal model is*

$$\mathcal{D}^{\mathcal{M}_{1,v_0}} = \{(v_0, 1), (v_0, 2), (v_0, 3), (v_0, 4), (1, 2, 3, 4)\}.$$

This is not a discrete graphical model since a graphical model would also include the interactions $(v_0, 1, 2), (v_0, 2, 3), (v_0, 3, 4), (v_0, 1, 4), (v_0, 1, 2, 3, 4)$. It was therefore crucial to set up our problem, as we did it in Section 2, within the framework of hierarchical loglinear models rather than the more restrictive class of discrete graphical models.

4.4 The composite maximum likelihood estimates

To obtain the composite mle of the global loglinear parameter θ, we do the following. First, for each $v \in V$, we compute the maximum likelihood estimates from L^{PS_v} as in (4.2) (in the conditional composite likelihood case) or for $L^{\mathcal{M}_{l,v}}$ in (4.12) (in the marginal composite likelihood case). Second, in each case, we retain only the estimate of θ^{PS_v}. Third, the global maximal composite likelihood estimate of each $\theta_j, j \in J$ is derived either by simple averaging of the different estimates of $\theta_j^{PS_v}$ or $\theta_j^{\mathcal{M}_{l,v}}, j \in J^{PS_v}$ obtained from the various local models or by more sophisticated ways such as described in Liu and Ihler (2012). This includes linear consensus, maximum consensus or ADMM. If one uses the two-hop neighbourhood, then the accuracy is such that usually simple averaging is sufficient to obtain very good accuracy.

If follows immediately that if we can prove that the mle of θ^{PS_v} obtained from L^{PS_v} and from $L^{M_{l,v}}$ are identical, then the maximal composite likelihood estimates of θ obtained from the conditional or the marginal composite likelihood by consensus will be the same. In the next subsection, we show that this is indeed the case.

4.5 Equality of the maximal conditional and marginal composite likelihood estimate

Let $\hat{\theta}^{\mathcal{M}_{l,v}}, l = 1, 2$ denote the maximum likelihood estimate of $\theta^{\mathcal{M}_{l,v}}$ obtained from the local likelihood (4.12).

Theorem 4.1 *The PS component of $\hat{\theta}^{\mathcal{M}_{1,v}}$,i.e. $(\hat{\theta}_j^{\mathcal{M}_{1,v}}, j \in J^{PS_v})$ is equal to the maximum likelihood estimate of θ^{PS_v} obtained from the local conditional likelihood (4.2).*
Similarly, The $2PS$ component of $\hat{\theta}^{\mathcal{M}_{2,v}}$,i.e. $(\hat{\theta}_j^{\mathcal{M}_{2,v}}, j \in J^{PS_{2,v}})$ is equal to the maximum likelihood estimate of $\theta^{PS_{2,v}}$ obtained from the local conditional likelihood (4.6).

The proof is given in the Appendix.

At this point, we ought to make an important observation. In the case of the two-hop marginal likelihood, it may happen that the buffer $\mathcal{B}_v$ is no longer equal to $\mathcal{N}_{2v}$. For example, if we consider a four-neighbour 5×10 lattice, the vertex 39 is such that $\mathcal{N}_v^2 = \{19, 28, 30, 37, 48, 50\}$ while $\mathcal{B}_v = \mathcal{N}_{2v} \backslash \{50\}$. The argument in the proof of Theorem 4.1 for j such that $S(j) \not\subset \mathcal{N}_{2v}$ then breaks down since in the $\mathcal{M}_{2,v}$-marginal model, some cells such as $i_{\mathcal{M}_v} = (i_{30} = 1, i_{50} = 1, 0_{\mathcal{M}_v \backslash \{30,50\}})$ with support in $\mathcal{N}_{2v}$ no longer have a complete support. This situation is illustrated in Figure 1 where for the sake of comparison, we also draw a vertex for which $\mathcal{N}_{2v} = \mathcal{B}_v$ and Theorem 4.1 applies..

In Tables 1 and 2, we give the numerical values of the maximum likelihood estimate $\theta_j, j \in J^{\mathcal{M}_{2,v}}$ obtained by the four local model $PS, PS_2, \mathcal{M}_{1,v}$ and $\mathcal{M}_{2,v}$ for j such that

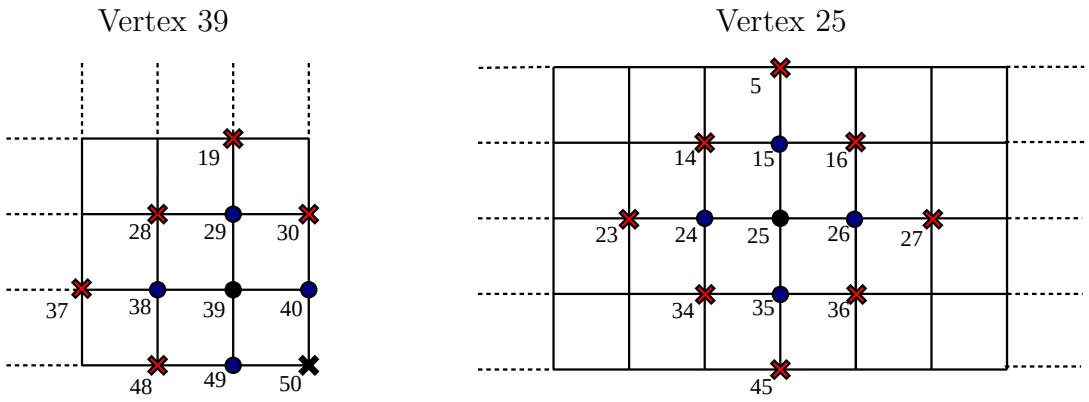

Figure 1: Two vertices in a 5×10 lattice: Theorem 4.1 applies for vertex 25 while it does not apply for vertex 39.

$j \in J^{PS_{25}}$ and for j such that $j \in J^{PS_{39}}$, respectively. We see that in the first case, the values of $\hat{\theta}_j$ obtained from the local likelihoods $l^{PS_{25}}$ and $l^{\mathcal{M}_{1,25}}$ are identical and similarly for those obtained from $l^{PS_{2,25}}$ and $l^{\mathcal{M}_{2,25}}$, while in the second case, the values obtained through the PS_2 and $\mathcal{M}_{2,v}$ models are slightly different. The values obtained from the PS and $\mathcal{M}_{1,v}$ models are identical since then $\mathcal{B}_v = \mathcal{N}_v$ and the proof of Theorem 4.1 does not break down.

Models	$\hat{\theta}_{25}$	$\hat{\theta}_{15,25}$	$\hat{\theta}_{24,25}$	$\hat{\theta}_{25,26}$	$\hat{\theta}_{25,35}$
$\mathcal{M}_{1,v}$	-0.0536	0.5914	-0.4808	-0.8314	-0.8461
$\mathcal{M}_{2,v}$	-0.0779	0.5221	-0.5310	-0.7274	-0.7459
(v, PS)	-0.0536	0.5914	-0.4808	-0.8314	-0.8461
$(v, 2PS)$	-0.0779	0.5221	-0.5310	-0.7274	-0.7459

Table 1: The composite mle of some $\theta_j, j \in J^{25,PS}$ in the 5×10 lattice

Remark 4.4 *The equality of the estimates holds also for the marginal estimates obtained by Mizrahi et al. (2014b) if, for q a clique of G and $v \in q \subset \mathcal{A}_q$, satisfying the strong LAP condition with respect to $\mathcal{A}_q$, we retain only the parameters $\theta_j, j \in J^{PS_v} \cap q$. We also note that Theorem 9 in that paper may not be verified in some cases. For example, take vertex 7 in a 3×3 lattice numbered from left to right starting with the top row, take $q = \{7, 8\}$*

Models	$\hat{\theta}_{39}$	$\hat{\theta}_{29,39}$	$\hat{\theta}_{38,39}$	$\hat{\theta}_{39,40}$	$\hat{\theta}_{39,49}$
$\mathcal{M}_{1,v}$	-1.0799	-0.3306	-0.3647	-0.5791	1.1749
$\mathcal{M}_{2,v}$	-1.0386	-0.3519	-0.5020	-0.5445	1.1946
(v, PS)	-1.0799	-0.3306	-0.3647	-0.5791	1.1749
$(v, 2PS)$	-1.0381	-0.3531	-0.5019	-0.5448	1.1947

Table 2: The composite mle of some $\theta_j, j \in J^{39,PS}$ in the 5×10 lattice

as the clique of interest. Then $\mathcal{A}_q = \{4, 7, 8\}$ *satisfies the strong LAP condition but* θ_8 *in the* $\mathcal{A}_q$*-marginal model cannot be equal to* θ_8 *in the joint model as our Lemma 4.2 shows.*

4.6 Existence of the mle and the composite likelihood estimates

To finish this section, we describe a numerical experiment illustrating the impact of the non existence of the global mle on the computations of the various composite maximum likelihood estimates.

For each sample size $N = 40, 60$ and 80, we consider two experiments, one where the data point $t = (t_j, j \in J) = (n(j_{S(j)}), j \in J)$ belongs to a face of the global model and one where it does not. For each experiment, we compute five estimates, the global mle, and the four composite likelihood estimates based on $\hat{\theta}^{\mathcal{M}_{1,v}}, \hat{\theta}^{PS_v}, \hat{\theta}^{\mathcal{M}_{2,v}}, \hat{\theta}^{PS_{2,v}}, v \in V$. For each one of the five estimates and for each experiment, we report the relative mean square error (MSE)

$$\frac{||\hat{\theta} - \theta^*||^2}{||\theta^*||^2}$$

where θ^* denotes the true value of the parameter. The results are given in Table 4.6

Sample size	40(on face)	40 (not on face)	60(on face)	60(not on face)	80(on face)	80(not on fac
Global MLE	102.72	3.1270	39.68	2.2620	27.12	1.5717
$\mathcal{M}_1$ MLE	134.94	3.6335	66.00	2.3214	43.46	1.6454
PS_1 MLE	127.64	3.6335	32.63	2.3214	24.34	1.6454
$\mathcal{M}_2$ MLE	340.81	3.1328	65.01	2.2700	42.67	1.5728
PS_2 MLE	84.52	3.1320	44.55	2.2709	29.76	1.5727

Table 3: The relative MSE of different estimates of θ for the 5×10 lattice when the mle exists and when it does not.

We note that when the global mle does not exist the mean square error for all experiments is much larger than when the mle exists indicating that some of the local mle estimates do not exist. When a local mle does not exist, a routine maximization of the local likelihood may lead to erroneous results. We have thus illustrated that the numerical results from each local maximization must be examined carefully to detect any potential existence problem. We note also that the data vector may be on a face of a global model without being on the face of any local model. In this case, clearly, the composite mle of θ is not a consistent estimate of θ_0.

5 Asymptotic properties of the maximum composite likelihood estimate

The asymptotic properties of the maximum composite likelihood estimate when p is fixed and N goes to $+\infty$ are well-known (see Jordan and Liang (2009)). In this section, we consider the asymptotic properties of the conditional composite mle (which is also the marginal composite mle) when both p and N go to $+\infty$. In Theorem 5.1 below, we give its rate of convergence to the true value θ^*. In order to compare the behaviour of the composite mle with the global mle, we also give, in Theorem 5.2, the rate of convergence of the global mle under the same asymptotic regime. It will be convenient to introduce the notation

$$f_j(x) = \prod_{l \in S(j)} \mathbb{1}(x_l = j_l) = \begin{cases} 1 & \text{if } j \lhd x \\ 0 & \text{otherwise} \end{cases},$$

and to write (4.3) as

$$p(x_v|x_{N_v}) = \frac{\exp\{\sum_{j \in J^{PS_v}} \theta_j f_j(x_v, x_{N_v})\}}{1 + \sum_{y_v \in I_v \setminus \{0\}} \exp\{\sum_{j \in J^{PS_v}} \theta_j f_j(y_v, x_{N_v})\}} . \tag{5.1}$$

In this section, we work exclusively with $l^{PS_v}(\theta^{PS_v})$. Therefore for simplicity of notation we write θ for θ^{PS_v}. Also, for convenience, we scale the log likelihood by the factor $\frac{1}{N}$. Then the v-local conditional log-likelihood function is

$$\begin{aligned} l^{PS_v}(\theta) &= \tfrac{1}{N} \textstyle\sum_{n=1}^N \log p(x_v^{(n)}|x_{N_v}^{(n)}) \\ &= \textstyle\sum_{j \in J^{PS_v}} \theta_j \tfrac{1}{N} \sum_{n=1}^N f_j(x_v^{(n)}, x_{N_v}^{(n)}) \\ &- \tfrac{1}{N} \textstyle\sum_{n=1}^N \log\{1 + \sum_{y_v \in I_v \setminus \{0\}} \exp\{\sum_{j \in J^{PS_v}} \theta_j f_j(y_v, x_{N_v}^{(n)})\}\} \end{aligned}$$

The sufficient statistic is $t_j = \frac{1}{N} \sum_{n=1}^N f_j(x_v^{(n)}, x_{N_v}^{(n)})$. We write

$$t_{J^{PS_v}} = [t_1, t_2, \cdots, t_{d_v}] \tag{5.2}$$

and

$$k^{PS_v}(\theta) = \frac{1}{N}\sum_{n=1}^{N} \log\{1 + \sum_{y_v \in I_v \setminus \{0\}} \exp\{\sum_{j \in J^{PS_v}} \theta_j f_j(y_v, x_{N_v}^{(n)})\}\} = \frac{1}{N}\sum_{n=1}^{N} \log Z^{n,v}(\theta),$$

where

$$Z^{n,v}(\theta) = 1 + \sum_{y_v \in I_v \setminus \{0\}} \exp\{\sum_{j \in J^{PS_v}} \theta_j f_j(y_v, x_{N_v}^{(n)})\}.$$

Then the log -likelihood function is

$$l^{PS_v}(\theta) = \sum_{j \in J^{PS_v}} \theta_j t_j - k^{PS_v}(\theta) \ .$$

Its first derivative is

$$\begin{aligned}
\frac{\partial l^{PS_v}(\theta)}{\partial \theta_k} &= t_k - \frac{\partial k^{PS_v}(\theta)}{\partial \theta_k}, \\
\frac{\partial k^{PS_v}(\theta)}{\partial \theta_k} &= \frac{1}{N}\sum_{n=1}^{N} \frac{\exp\{\sum_{j \in J^{PS_v}} \theta_j f_j(k_v, x_{N_v}^{(n)})\}}{Z^{n,v}(\theta)} f_k(k_v, x_{N_v}^{(n)})
\end{aligned}$$

with

$$\frac{\exp\{\sum_{j \in J^{PS_v}} \theta_j f_j(k_v, x_{N_v}^{(n)})\}}{Z^{n,v}(\theta)} = p(X_v = k_v | x_{N_v}^{(n)}) \tag{5.3}$$

We now want to compute $\frac{\partial^2 l^{PS_v}(\theta)}{\partial \theta_k \partial \theta_l} = -\frac{\partial^2 k^{PS_v}(\theta)}{\partial \theta_k \partial \theta_l}$, $k, l \in J^{PS_v}$. To simplify further our notation, we set

$$z_{y_v}(\theta) = \sum_{j \in J^{PS_v}} \theta_j f_j(y_v, x_{N_v}^{(n)}). \tag{5.4}$$

For $k_v = l_v$, using (5.3), we obtain

$$\begin{aligned}
\frac{\partial^2 k^{PS_v}(\theta)}{\partial \theta_k \partial \theta_l} &= \frac{1}{N}\sum_{n=1}^{N} \left(\frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)} - (\frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)})^2 \right) f_k(k_v, x_{N_v}^{(n)}) f_l(l_v, x_{N_v}^{(n)}) \\
&= \frac{1}{N}\sum_{n=1}^{N} \left(p(X_v = k_v | x_{N_v}^{(n)}) - p(X_v = k_v | x_{N_v}^{(n)})^2 \right) f_k(k_v, x_{N_v}^{(n)}) f_l(l_v, x_{N_v}^{(n)}) \ .
\end{aligned}$$

if $k_v \neq l_v$, then

$$\begin{aligned}
\frac{\partial^2 k^{PS_v}(\theta)}{\partial \theta_k \partial \theta_l} &= \frac{1}{N}\sum_{n=1}^{N} -\frac{\exp z_{k_v}(\theta) \exp z_{l_v}(\theta)}{(Z^{n,v}(\theta))^2} f_k(k_v, x_{N_v}^{(n)}) f_l(l_v, x_{N_v}^{(n)}) \\
&= \frac{1}{N}\sum_{n=1}^{N} (-p(X_v = k_v | x_{N_v}^{(n)}) p(X_v = l_v | x_{N_v}^{(n)})) f_k(k_v, x_{N_v}^{(n)}) f_l(l_v, x_{N_v}^{(n)}) \ .
\end{aligned}$$

Let $W^{n,v} = (f_j(j_v, x_{N_v}^{(n)}), j \in J^{PS_v})$ be the $d_v \times 1$ vector of indicators. We introduce the notation

$$\eta_{k,l}^{n,v}(\theta, x_{N_v}^{(n)}) = \begin{cases} \frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)} - (\frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)})^2, & \text{if } k_v = l_v \\ -\frac{\exp z_{k_v}(\theta) \exp z_{l_v}(\theta)}{(Z^{n,v}(\theta))^2}, & \text{if } k_v \neq l_v \ . \end{cases} \tag{5.5}$$

Let $H^{n,v}(\theta, x_{N_v}^{(n)})$ be the $d_v \times d_v$ matrix with (k,l) entry $\eta_{k,l}^{n,v}(\theta, x_{N_v}^{(n)})$. Then the Fisher information matrix derived from l^{PS_v} is

$$(k^{PS_v})''(\theta) = \frac{1}{N} \sum_{n=1}^{N} H^{n,v}(\theta, x_{N_v}^{(n)}) \circ [W^{n,v}(W^{n,v})^t]$$

where $\circ$ denotes the Hadamard product of two matrices. We make two assumptions on the behaviour of the cumulant generating function $k^{PS_v}, v \in V$ at θ^*, similar to those made by Ravikumar et al. (2010) and Meng (2014):

(A) For the design matrix of the v-local conditional models, we assume that there exists $D_{max} > 0$ such that

$$\max_{v \in V} \lambda_{max}\Big(\frac{1}{N} \sum_{n=1}^{N} W^{n,v}(W^{n,v})^t\Big) \leq D_{max};$$

(B) We assume the minimum eigenvalue of the Fisher Information matrices $(k^{PS_v})''(\theta^*),\ v \in V$ is bounded, i.e., there exists $C_{min} > 0$ such that

$$C_{min} = \min_{v \in V} \lambda_{min} \frac{1}{N} \sum_{n=1}^{N} \big[H^{n,v}(\theta^*, x_{N_v}^{(n)}) \circ [W^{n,v}(W^{n,v})^t]\big].$$

We are now ready to state our theorem on the asymptotic behaviour of the composite mle $\hat{\theta}$ obtained by averaging the mle of $\hat{\theta}_j^{PS_v}$ obtained from the various v-neighbourhood as indicated in Section 4.4.

Theorem 5.1 *Assume conditions (A) and (B) hold. If the sample size N and $|V| = p$ satisfy*

$$\frac{N}{\log p} \geq \max_{v \in V} (\frac{10 C D_{max} d_v}{C_{min}^2})^2,$$

where C is a positive constant such that $p^{\frac{C^2}{2}} \geq 2 \sum_{v \in V} d_v$, then the conditional composite mle $\hat{\theta} = (\hat{\theta}_j, j \in J)$ is such that

$$\|\hat{\theta} - \theta^*\|_F \leq \frac{5C}{C_{min}} \sqrt{\frac{\sum_{v \in V} d_v \log p}{N}} \tag{5.6}$$

with probability greater than $1 - \frac{2\sum_{v\in V} d_v}{p^{\frac{C^2}{2}}}$.

The proof is given in the Appendix. With a similar argument, we can derive the behaviour of the global mle, which we will denote by $\hat{\theta}^G$. We need to make assumptions similar to (A) and (B). We assume that

$$(A') \;\text{ there exists }\; D_{max} > 0 \;\text{ such that }\; \lambda_{max}\Big(\sum_{i\in I} f_i \otimes f_i\Big) \leq D_{max},$$

$$(B') \;\; 0 < \kappa^* = \lambda_{min}\Big[k''(\theta^*)\Big].$$

The asymptotic behaviour of $\hat{\theta}^G$ is given in the following theorem.

Theorem 5.2 *Assume conditions* (A') *and* (B') *hold. If* N *and* p *satisfy the condition*

$$\frac{N}{\log p} \geq (\frac{40C|J|D_{max}}{\kappa^{*2}})^2,$$

where C *is a positive constant such that* $p^{2C^2} \geq 2|J|$*, then the global mle* $\hat{\theta}^G = (\hat{\theta}^G_j, j \in J)$ *is such that*

$$\|\hat{\theta}^G - \theta^*\|_F \leq \frac{5C}{\kappa^*}\sqrt{\frac{|J|\;\log p}{N}} \tag{5.7}$$

with probability greater than $1 - \frac{2|J|}{p^{2C^2}}$.

The proof is provided in the Supplementary file. Comparing Theorem 5.1 and 5.2, we see that for $\frac{N}{p} = \mathcal{O}(|J|^2)$, $\|\hat{\theta}^G - \theta^*\|_F = \mathcal{O}(\sqrt{\frac{|J|\;\log p}{N}})$ with high probability while for $\frac{N}{p} = \mathcal{O}(\max_{v\in V}(d_v^2))$, $\|\hat{\theta} - \theta^*\|_F = \mathcal{O}(\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}})$. This implies that for the composite mle, the requirement on the sample size N are not as stringent as for the global mle but of course, we lose some accuracy in the approximation of θ^*. The situation is, however, not bad since

$$\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}} \Big/ \sqrt{\frac{|J|\;\log p}{N}} = \sqrt{\frac{\sum_{v\in V} d_v}{|J|}}$$

which is the square root of the ratio of the sum over $v \in V$ of the number of parameters in the v-local conditional models and the number of parameters in the global model. If the number of neighbours for each vertex is bounded by d, we see that this ratio is at most equal to $\frac{2^{d+1}}{|J|}$ and usually much smaller than that. For example, in an Ising model, $|J| = p + |E|$ and $\sum_{v\in V} d_v = p + 2|E|$ and therefore $\frac{\sum_{v\in V} d_v}{|J|} = 1 + \frac{|E|}{p+|E|} \leq 2$.

6 Conclusion

In this paper, we have taken a local approach to study the existence of the maximum likelihood estimate of the canonical parameter θ in a high-dimensional discrete graphical model. We have shown that we can use smaller graphical models to detect the nonexistence of this mle. We have also taken a local approach to the estimation of θ by looking at various possible versions of the composite likelihood estimate, based on local conditional or marginal likelihoods and we have shown that the two approaches yield the same estimate of θ. Through a numerical experiment, we have illustrated how we can be led to an incorrect maximum composite likelihood estimate of θ if the global estimate does not exist. Finally, we have described the asymptotic behaviour of the maximum composite likelihood estimate of θ when both the dimension p of the model and the sample size N tend to infinity. We have shown that when the number of neighbours of each $v \in V$ is bounded, the rate of convergence of the composite mle is comparable to that of the global mle.

References

Besag, J., (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. Roy. Statist. Soc., Ser. B*, **36**, 192-236.

Birch, M.W., (1963), Maximum likelihood in three-way contingency tables, *J. Roy. Statist. Soc., Ser. B*, **25**, 220-233.

Eriksson, N., Fienberg, S. E., Rinaldo, A. and Sullivant, S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J. Symbolic Comput.*, **41**, 222-233.

Fienberg, S. E. and Rinaldo, A. (2007). Three centuries of categorical data analysis: Loglinear models and maximum likelihood estimation. *J. Statist. Plann. Inference*, **137**, 3430-3445.

Fienberg, S. E. and Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.*, **40**, 9961023.

Fienberg, S. E. and Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. Technical report, Carnegie Mellon Univ. Available at http://www.stat.cmu.edu/ arinaldo/Fienberg-Rinaldo-Supplementary-Material.pdf.

Geyer, C.J. (1991). Markov chain Monte-Carlo maximum likelihood, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156-163.

Haberman, S.J., (1974). The Analysis of Frequency Data. Univ. Chicago Press, Chicago.

Hoeffding, W.(1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13-30.

Jordan, M. I. , Ghahramani, Z., Jaakkola, T. S. & Saul, L. K.(1999). An introduction to variational methods for graphical models. *Machine Learning*, **37**,183-233.

Lauritzen, S.L. (1996). *Graphical Models*, Oxford Science Publications.

Letac, G. and Massam, H., (2012), Bayes regularization and the geometry of discrete hierarchical loglinear models, *Ann. Statist.*, **40**, 861-890.

Liang, P. and Jordan, M.I., (2008), An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators, *Proceedings of the 25th International Conference on Machine Learning*.

Lindsay, B. G. (1988). Composite likelihood methods, *Contemp. Math.*, **80**, 221-239.

Liu, Q. and Ihler, A., (2012), Distributed parameter estimation via pseudo-likelihood, *International Conference on Machine Learning, (ICML)*.

Meng, Z., Wei, D. Wiesel, A. and Hero, A.O. III, (2013), Distributed learning of Gaussian graphical models via marginal likelihood, *J. Mach. Learn. Res. W & CP*, **31**, 39-47.

Meng, D. Wei, A. Wiesel, A. Hero, (2014) "Distributed Learning of Gaussian Graphical Models via Marginal Likelihoods," *IEEE Trans. Signal Process.*, **62**, 5425-5438.

Mizrahi, Y.D., Denil, M. and de Freitas, N., (2014a), Distributed Parameter Estimation in Probabilistic Graphical Models, *Arxiv, 1406.3070*.

Mizrahi, Y.D., Denil, M. and de Freitas, N., (2014b), Distributed Parameter Estimation in Probabilistic Graphical Models, *Advances in Neural Information Processing Systems (NIPS)*.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional graphical Ising model selection using l_1-regularized logistic regression, *Ann. Statist.*, **38**, 1287-1319.

Schmidt, M.W., Berg, E., Friedlander, M.P. and Murphy, K.P., (2009), Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm, *J. Mach. Learn. Res., W&CP 5*, 73-80.

Snijders, T.A.B., (2002). Markov chain Monte Carlo estimation of exponential random graph models, *J. of Social Structure*, **3**, 1-40.

Wainwright, M. and Jordan, M.I., (2008). Graphical Models, Exponential Families, and Variational Inference, *Found. Trends Mach. Learn.*, **1**, 1-305.

Wiesel, A. and Hero, A.O. III, (2012), Distributive covariance estimation in Gaussian graphical models, *IEEE Trans. signal process.*, **60**, 211-220.

7 Appendix

7.1 Proof of Lemma 4.1

We will use the notation $j \lhd_0 j'$ to mean that $j \lhd j'$ or $j = 0$, the zero cell. Let $p^{\mathcal{M}_v}(i)$ denote the marginal probability of $i \in I_{\mathcal{M}_v}$. We know that the $\mathcal{M}_v$-marginal distribution of $X_{\mathcal{M}_v}$ is multinomial. By the general parametrization of the multinomial model (2.2), for $j \in J, S(j) \subset \mathcal{M}_v$, since $S(j)$ is complete,

$$\theta_j^{\mathcal{M}_v} = \sum_{j' \in J,\, j' \lhd j} (-1)^{|S(j)|-|S(j')|} \log \frac{p^{\mathcal{M}_v}(j')}{p^{\mathcal{M}_v}(0)}, \tag{7.1}$$

where by abuse of notation, j such that $S(j) \subset \mathcal{M}_v$ is considered as an element of $I_{\mathcal{M}_v}$.

Moreover,

$$\begin{aligned} p^{\mathcal{M}_v}(j) &= \sum_{i \in \mathcal{I}:\, i_{\mathcal{M}_v} = j} p(i) = \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v} = j} \exp\{ \sum_{j' \mid j' \lhd_0 j} \theta_{j'} + \sum_{\substack{j' \mid j' \lhd i \\ j' \not\lhd j \\ j'_{\mathcal{M}_v} \lhd_0 j}} \theta_{j'} \} \\ &= \Big(\exp \sum_{j' \mid j' \lhd_0 j} \theta_{j'} \Big) \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v} = j} \exp \sum_{\substack{j' \mid j' \lhd i \\ j' \not\lhd j \\ j'_{\mathcal{M}_v} \lhd_0 j}} \theta_{j'} \Big) . \end{aligned}$$

Therefore $\log p^{\mathcal{M}_v}(j) = \sum_{j' \mid j' \lhd_0 j} \theta_{j'} + \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v} = j} \exp \sum_{\substack{j' \mid j' \lhd i \\ j' \not\lhd j}} \theta_{j'} \Big)$, which we can write

$$\sum_{j' \mid j' \lhd_0 j} \theta_{j'} = \log p^{\mathcal{M}_v}(j) - \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v} = j} \exp \sum_{\substack{k \mid k \lhd i \\ k \not\lhd j}} \theta_k \Big) . \tag{7.2}$$

Moebius inversion formula states that for $a \subseteq V$ an equality of the form $\sum_{b \subseteq a} \Phi(b) = \Psi(a)$ is equivalent to $\Phi(a) = \sum_{b \subseteq a} (-1)^{|a \setminus b|} \Psi(b)$. Here, using a generalization of the Moebius inversion formula to the partially ordered set given by $\lhd$ on J, we derive from (7.2) that for $j \in J^{\mathcal{M}_v} \subset J$

$$\begin{aligned}
\theta_j &= \sum_{j' \mid j' \lhd_0 j} (-1)^{|S(j)-S(j')|} \log p^{\mathcal{M}_v}(j') \\
&\quad - \sum_{j' \mid j' \lhd_0 j} (-1)^{|S(j)-S(j')|} \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v} = j'} \exp \sum_{\substack{k \mid k \lhd i \\ k \not\lhd j'}} \theta_k \Big) \\
&= \theta_j^{\mathcal{M}_v} - \sum_{j' \mid j' \lhd_0 j} (-1)^{|S(j)-S(j')|} \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v} = j'} \exp \sum_{\substack{k \mid k \lhd i \\ k \not\lhd j'}} \theta_k \Big) \qquad (7.3)
\end{aligned}$$

which we prefer to write as (4.9).

7.2 Proof of Lemma 4.2

Since (4.9) is already proved, *(2.)* holds. Let us prove that *(1.)* holds, i.e., that when $S(j) \not\subset \mathcal{B}_v$, the alternating sum on the right-hand side of (4.9) is equal to 0. Since $j \in J$, $S(j)$ is necessarily complete and $j' \lhd j$ is obtained by removing one or more vertices from $S(j)$.

If $S(j) \cap \mathcal{B}_v \neq \emptyset$ but $S(j) \not\subset \mathcal{B}_v$, there is at least one vertex $w \in S(j)$ which is not in $\mathcal{B}_v$. Let l_0 and l_w be the log terms in the alternating sum corresponding to $j' = 0$ and $j'_w \lhd j$ such that $S(j'_w) = \{w\}$ respectively. Since for any neighbours u of w in $\mathcal{M}_v$ and for any $i \in I$ such that $i_{\mathcal{M}_v} = j'$, the u-th coordinate i_u must be zero and since w cannot have a neighbour outside $\mathcal{M}_v$, the set $\{\theta_k, k \lhd i^{(1)}, k \not\lhd j'\}$ in l_0 for $i^{(1)}$ such that $i^{(1)}_{\mathcal{M}_v} = 0$ is the same as the set $\{\theta_k, k \lhd i^{(2)}, k \not\lhd j'\}$ in l_w for $i^{(2)}$ such that $i^{(2)}_{\mathcal{M}_v} = j'_w$ and $i^{(2)}_{V \setminus \mathcal{M}_v} = i^{(1)}_{V \setminus \mathcal{M}_v}$. The terms in l_0 and l_w in (4.9) are therefore exactly the same except for their sign and these two terms cancel out. Similarly, for any given $j' \lhd j$ with $w \notin S(j')$, let $j'_w \in J$ be such that $S(j'_w) = S(j) \cup \{w\}$ and $j'_w \lhd j$, then, the set $\theta_k, k \lhd i^{(1)}, k \not\lhd j'$ in $l_{j'}$ and the set $\theta_k, k \lhd i^{(2)}, k \not\lhd j'_w$ in $l_{j'_w}$ are identical where, similarly to the argument above, $i^{(1)}$ is such that $i^{(1)}_{\mathcal{M}_v} = j'$ and $i^{(2)}$ is such that $i^{(1)}_{\mathcal{M}_v} = j'_w$ and $i^{(2)}_{V \setminus \mathcal{M}_v} = i^{(1)}_{V \setminus \mathcal{M}_v}$. Therefore the terms $l_{j'}$ and $l_{j'_w}$ cancel out and *(1.)* is proved.

To prove that *(3.)* holds, following (2.2), we have, for $S(i) = E \subset \mathcal{M}_v$

$$\begin{aligned}
\theta_i^{\mathcal{M}_v} &= \sum_{F\subset E}(-1)^{|E\setminus F|}\log p^{\mathcal{M}_v}(i_F, 0_{\mathcal{M}_v\setminus F}) \\
&= \sum_{F\subset E}(-1)^{|E\setminus F|}\log\Big(p(i_F, 0_{V\setminus F}) + \sum_{L\subset V\setminus\mathcal{M}_v}\sum_{k_L\in I_L} p(i_F, 0_{\mathcal{M}_v\setminus F}, k_L, 0_{V\setminus(\mathcal{M}_v\cup L)})\Big) \\
&= \sum_{F\subset E}(-1)^{|E\setminus F|}\log\Big(\exp(\sum_{j\in J, j\lhd i_F}\theta_j) + \sum_{L\subset V\setminus F}\sum_{k_L\in I_L}\exp(\sum_{j\in J, j\lhd i_F}\theta_j + \sum_{j\not\lhd i_F, j\lhd(i_F,k_L)}\theta_j)\Big) \\
&= \sum_{F\subset E}(-1)^{|E\setminus F|}\log\Big(\exp(\sum_{j\in J, j\lhd i_F}\theta_j)\Big) \qquad (7.4)\\
&\quad + \sum_{F\subset E}(-1)^{|E\setminus F|}\log(1 + \sum_{L\subset V\setminus F}\sum_{k_L\in I_L}\exp(\sum_{j\not\lhd i_F, j\lhd(i_F,k_L)}\theta_j)\Big) \\
&= \theta_i + \sum_{F\subset E}(-1)^{|E\setminus F|}\log(1 + \sum_{L\subset V\setminus F}\sum_{k_L\in I_L}\exp(\sum_{j\not\lhd i_F, j\lhd(i_F,k_L)}\theta_j)\Big) \qquad (7.5)
\end{aligned}$$

Now, following an argument similar to that of *(1.)* above, we can show that the second component of the sum in (7.5) is equal to zero. It follows that when $\theta_i = 0$ then $\theta_i^{\mathcal{M}_v} = 0$. This completes the proof of Lemma 4.2.

7.3 Proof of Theorem 4.1

The local relaxed marginal loglikelihood is

$$\begin{aligned}
l^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}}) &= \sum_{k=1}^{N}\log p^{\mathcal{M}_{l,v}}(X_{\mathcal{M}_v} = i^{(k)}_{\mathcal{M}_v}) = \sum_{i_{\mathcal{M}_v}\in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v})\log p^{\mathcal{M}_{l,v}}(i_{\mathcal{M}_v}) \\
&= \langle\theta^{\mathcal{M}_{l,v}}, t^{\mathcal{M}_{l,v}}\rangle - Nk^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})
\end{aligned}$$

It is immediate to see that $\frac{\partial l^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})}{\partial\theta_j} = t(j) - p^{\mathcal{M}_{l,v}}(j_{S(j)})$ where $p^{\mathcal{M}_{l,v}}(j_{S(j)})$ denotes the $j_{S(j)}$-marginal cell probability in the $\mathcal{M}_{l,v}$-marginal model. Therefore the likelihood equations $\frac{\partial l^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})}{\partial\theta_j} = 0,\ j \in J^{\mathcal{M}_{l,v}}$ yield

$$t(j) - p^{\mathcal{M}_{l,v}}(j_{S(j)}) = 0, \qquad (7.6)$$

where $t(j) = n(j_{S(j)})$.

For the argument to follow is essentially the same for the one-hop or two-hop neighbourhood. We present it for the more general case of the two hop neighbourhood. The local conditional log likelihood is

$$\begin{aligned}
l^{v,2PS}(\theta^{v,2PS}) &= \sum_{i_{\mathcal{M}_v} \in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log \frac{p(X_v = i_v, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}, X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})}{p(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})} \\
&= \sum_{i_{\mathcal{M}_v} \in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log \frac{p^{\mathcal{M}^{2,v}}(X_{\mathcal{M}_v} = i_{\mathcal{M}_v})}{p^{\mathcal{M}^{2,v}}(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})} \\
&= \sum_{i_{\mathcal{M}_v} \in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log p^{\mathcal{M}^{2,v}}(X_{\mathcal{M}_v} = i_{\mathcal{N}_v}) - \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log p^{\mathcal{M}^{2,v}}(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}) \\
&= l^{\mathcal{M}_{2,v}}(\theta^{\mathcal{M}_{2,v}}) - \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log \sum_{x_{v \cup \mathcal{N}_v} \in I_{v \cup \mathcal{N}_v}} p^{\mathcal{M}^{2,v}}(X_{v \cup \mathcal{N}_v} = x_{v \cup \mathcal{N}_v}, X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}) \\
&= l^{\mathcal{M}_{2,v}}(\theta^{\mathcal{M}_{2,v}}) - Q \qquad (7.7)
\end{aligned}$$

where

$$Q = \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log \sum_{x_{v \cup \mathcal{N}_v} \in I_{v \cup \mathcal{N}_v}} \exp\Big(\theta_0 + \sum_{\substack{k \lhd (x_{v \cup \mathcal{N}_v}, i_{\mathcal{N}_{2v}}) \\ k \in J^{\mathcal{M}_{2,v}}}} \theta_k\Big) \qquad (7.8)$$

and $\theta_0 = -\log(\sum_{i_{\mathcal{M}_v} \in I_{\mathcal{M}_v}} \exp \sum_{k \lhd i_{\mathcal{M}_v}, k \in J^{\mathcal{M}_{2,v}}} \theta_k)$. The second equality above is due to the fact that in the expression (4.3) of $\frac{p(X_v = i_v, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}, X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})}{p(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})}$, the θ_j such that $S(j) \notin \mathcal{M}_v$ and the θ_j such that $S(j) \subset \mathcal{N}_{2v}$ cancel out at the numerator and denominator and it therefore does not matter, for the conditional distribution of $X_{v \cup \mathcal{N}_v}$ given $X_{\mathcal{N}_{2v}}$, what the relationship between the neighbours are. The only thing that matters is the relationship between the vertices in $v \cup \mathcal{N}_v$ and the vertices in $\mathcal{M}_v$ and according to Lemma 4.2, that remains unchanged when we change from the global model to the $\mathcal{M}_{2,v}$-marginal models.

We will now differentiate the expression of $l^{v,2PS}$ in (7.8) with respect to $\theta_j, j \in J^{\mathcal{M}_{2,v}}$. We first note that

$$\frac{\partial \theta_0}{\partial \theta_j} = p^{\mathcal{M}^{2,v}}(j_{S(j)}).$$

If we use the notation

$$\mathbf{1}_{j \lhd (x_{v \cup \mathcal{N}_v}, i_{\mathcal{N}_{2v}})} = \begin{cases} 1 & \text{if } j \lhd (x_{v \cup \mathcal{N}_v}, i_{\mathcal{N}_{2v}}) \\ 0 & \text{otherwise} \end{cases},$$

and the notation $p^{\mathcal{M}_{2,v}}(i_E),\ E \subset \mathcal{M}_v$ to denote the marginal probability of $X_E = i_E$ in the $\mathcal{M}_{2,v}$-marginal model, we have

$$\frac{\partial Q}{\partial \theta_j} = \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \frac{\sum_{x_{v\cup\mathcal{N}_v} \in I_{v\cup\mathcal{N}_v}} p^{\mathcal{M}^{2,v}}(x_{v\cup\mathcal{N}_v}, i_{\mathcal{N}_{2,v}}) \Big(\mathbf{1}_{j \lhd (x_{v\cup\mathcal{N}_v}, i_{\mathcal{N}_{2v}})} - p^{\mathcal{M}^{2,v}}(j_{S(j)}) \Big)}{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2,v}})}.$$

If $j \in J^{\mathcal{M}_{2,v}}$ is such that $S(j) \subset \mathcal{N}_{2v}$, then $\mathbf{1}_{j \lhd (x_{v\cup\mathcal{N}_v}, i_{\mathcal{N}_{2v}})} = \mathbf{1}_{j_{\mathcal{N}_{2v}} \lhd i_{\mathcal{N}_{2v}}}$ and

$$\begin{aligned}
\frac{\partial Q}{\partial \theta_j} &= \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \frac{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}}) \Big(\mathbf{1}_{j_{\mathcal{N}_{2,v}} \lhd i_{\mathcal{N}_{2v}}} - p^{\mathcal{M}^{l,v}}(j_{S(j)}) \Big)}{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}})} \\
&= \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \Big(\mathbf{1}_{j_{\mathcal{N}_{2,v}} \lhd i_{\mathcal{N}_{2,v}}} - p^{\mathcal{M}^{2,v}}(j_{S(j)}) \Big) \\
&= n(j_{S(j)}) - N p^{\mathcal{M}^{2,v}}(j_{S(j)})
\end{aligned}$$

At the mle of the local $\mathcal{M}_{l,v}$ model, from standard likelihood equations (see Lauritzen, 1996, Theorem 4.11), we have $\hat{p}^{\mathcal{M}^{l,v}}(j_{S(j)}) = \frac{n(j_{S(j)})}{N}$ and therefore

$$\frac{\partial Q}{\partial \theta_j} = 0, \quad j \in J^{\mathcal{M}_{2,v}},\ S(j) \subset \mathcal{N}_{2v}. \tag{7.9}$$

If $j \in J^{\mathcal{M}_{2,v}}$ is such that $S(j) \not\subset \mathcal{N}_{2v}$, i.e. if $j \in J^{v,2PS}$,

$$\begin{aligned}
\frac{\partial Q}{\partial \theta_j} &= \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \frac{p^{\mathcal{M}^{2,v}}(j_{S(j)\cap(v\cup\mathcal{N}_v)}, i_{\mathcal{N}_{2v}}) \mathbf{1}_{j_{\mathcal{N}_{2v}} \lhd i_{\mathcal{N}_{2v}}} - p^{\mathcal{M}^{2,v}}(j_{S(j)}) p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}})}{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}})} \\
&= -p^{\mathcal{M}^{2,v}}(j_{S(j)}) \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) + \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} \frac{n(i_{\mathcal{N}_{2v}})}{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}})} p^{\mathcal{M}^{2,v}}(j_{S(j)\cap(v\cup\mathcal{N}_v)}, i_{\mathcal{N}_{2v}}) \mathbf{1}_{j_{\mathcal{N}_{2v}} \lhd i_{\mathcal{N}_{2v}}}
\end{aligned}$$

Since in the $\mathcal{M}_{2,v}$-marginal model, all the vertices in $\mathcal{N}_{2,v}$ are connected by construction, at the mle of the local $\mathcal{M}_{2,v}$ model, $\hat{p}^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}}) = \frac{n(i_{\mathcal{N}_{2v}})}{N}$ and therefore

$$\begin{aligned}
\frac{\partial Q}{\partial \theta_j} &= -N p^{\mathcal{M}^{2,v}}(j_{S(j)}) + N \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} p^{\mathcal{M}^{2,v}}(j_{S(j)\cap(v\cup\mathcal{N}_v)}, i_{\mathcal{N}_{2v}}) \mathbf{1}_{j_{\mathcal{N}_{2v}} \lhd i_{\mathcal{N}_{2v}}} \\
&= -N p^{\mathcal{M}^{2,v}}(j_{S(j)}) + N p^{\mathcal{M}^{2,v}}(j_{S(j)}) = 0
\end{aligned} \tag{7.10}$$

It follows from (7.9) and (7.10) that the $2PS$ component of $\hat{\theta}^{\mathcal{M}_{2,v}}$, i.e.

$$\hat{\theta}_j^{\mathcal{M}_{2,v}}, j \in J^{2,PS}$$

is the mle of the local two-hop conditional likelihood. We therefore have

$$\hat{\theta}^{v,2PS} = (\hat{\theta}^{\mathcal{M}_{2,v}})_{2PS}.$$

7.4 Proof of Theorem 5.1

To prove Theorem 5.1, we need two preliminary results.

Lemma 7.1 *Let $\theta^{v,*} = (\theta^*)^{PS_v}$ be the true value of the parameter for the conditional model of X_v given X_{N_v}, and let $\hat{\theta}^{PS_v}$ be the value of θ^{PS_v} that maximizes $l^{PS_v}(\theta)$. Then, for $t_{J^{PS_v}}$ as in (5.2), if there exists $\epsilon > 0$ such that*

$$\|t_{J^{PS_v}} - (k^{PS_v})'(\theta^{v,*})\|_\infty \leq \epsilon \leq \frac{C_{min}^2}{10 D_{max} d_v} \tag{7.11}$$

then

$$\|\hat{\theta}^{PS_v} - \theta^{v,*}\|_F \leq \frac{5\sqrt{d_v}\epsilon}{C_{min}} \tag{7.12}$$

Proof. To simplify our notation in this proof, we drop any subscripts and superscripts containing v or PS, except when it is necessary to keep them to make the argument clear.

Let $Q(\Delta) = l(\theta^*) - l(\theta^* + \Delta)$. Clearly $Q(0) = 0$ and $Q(\hat{\Delta}) \leq Q(0) = 0$, where $\hat{\Delta} = \hat{\theta} - \theta^*$. Let $||\Delta||_F = \sqrt{\sum_{j \in J^{PS_v}} \Delta_j^2}$ denote the Frobenius norm of Δ. Define $C(\delta) = \{\Delta | \quad \|\Delta\|_F = \delta\}$. Since $Q(\Delta)$ is a convex function of Δ, if we can prove

$$\inf_{\Delta \in C(\delta)} Q(\Delta) > 0, \tag{7.13}$$

then, by convexity of Q, it will follow that $\hat{\Delta}$ must lie within the sphere defined by $C(\delta)$, i.e. $\|\hat{\Delta}\|_F \leq \delta$. We are now going to prove that there exists $\delta > 0$ such that on $C(\delta)$, $Q(\Delta) > 0$. For $\Delta \in C(\delta)$, we have

$$\begin{aligned} Q(\Delta) &= l(\theta^*) - l(\theta^* + \Delta) = \theta^{*t} t - k(\theta^*) - ((\theta^* + \Delta)^t t - k(\theta^* + \Delta)) \\ &= k(\theta^* + \Delta) - k(\theta^*) - \Delta^t t = \Delta^t k'(\theta^*) + \tfrac{1}{2}\Delta^t k''(\theta^* + \alpha\Delta)\Delta - \Delta^t t, \quad \alpha \in [0,1] \\ &= \underbrace{\Delta^t [k'(\theta^*) - t]}_{Q_1} + \underbrace{\frac{1}{2}\Delta^t k''(\theta^* + \alpha\Delta)\Delta}_{Q_2} \end{aligned} \tag{7.14}$$

By Cauchy-Schwartz inequality, we have the following bound for Q_1.

$$|Q_1| = |\Delta^t[k^{'}(\theta^*) - t]| \quad \leq \quad \|k^{'}(\theta^*) - t\|_\infty ||\Delta||_1 \leq \epsilon\sqrt{d}\|\Delta\|_F = \epsilon\sqrt{d}\delta \tag{7.15}$$

For Q_2, we have

$$Q_2 \geq \frac{1}{2}\|\Delta\|_F^2 \min_{\alpha\in[0,1]} \lambda_{min} k^{''}(\theta^* + \alpha\Delta) = \frac{1}{2}\delta^2 \min_{\alpha\in[0,1]} \lambda_{min} k^{''}(\theta^* + \alpha\Delta) \tag{7.16}$$

We now want to bound the term $q = \min_{\alpha\in[0,1]} \lambda_{min}[k^{''}(\theta^*+\alpha\Delta)]$ from below. Following (5.4), we can write $z_{y_v}(\theta + \alpha\Delta) = \sum_{j\in J; v\in S(j)}(\theta_j + \alpha\Delta_j)f_j(y_v, x^{(n)}_{N_v})$, then we can rewrite the entries of H in (5.5) as

$$\eta^{n,v}_{k,l}(\theta^* + \alpha\Delta, x^{(n)}_{N_v}) = \begin{cases} \frac{\exp z_{k_v}(\theta^*+\alpha\Delta)}{1+\sum_{y_v\in I_v\setminus\{0\}} \exp z_{y_v}(\theta^*+\alpha\Delta)} - (\frac{\exp z_{k_v}(\theta^*+\alpha\Delta)}{1+\sum_{y_v\in I_v\setminus\{0\}} \exp z_{k_v}(\theta^*+\alpha\Delta)})^2, & \text{if } k_v = l_v \\ -\frac{\exp z_{k_v}(\theta^*+\alpha\Delta)\exp z_{l_v}(\theta^*+\alpha\Delta)}{(1+\sum_{y_v\in I_v\setminus\{0\}} \exp z_{y_v}(\theta^*+\alpha\Delta))^2}, & \text{if } k_v \neq l_v \end{cases}$$

then

$$\frac{\partial\eta^{n,v}_{k,l}(\theta^* + \alpha\Delta, x^{(n)}_{N_v})}{\partial\alpha} = \sum_{y_v\in I_v\setminus\{0\}} (\eta^{n,v}_{k,l})^{'}_{y_v}(\theta^* + \alpha\Delta, x^{(n)}_{N_v})\frac{\partial z_{y_v}}{\partial\alpha},$$

where $(\eta^{n,v}_{k,l})^{'}_{y_v}(\theta^* + \alpha\Delta, x^{(n)}_{N_v}) = \frac{\partial\eta^{n,v}_{k,l}(\theta^*+\alpha\Delta, x^{(n)}_{N_v})}{\partial z_{y_v}}$. It is easy to see that these derivatives can all be expressed in terms of probabilities of the type (5.3) and that they are always less than 1 in absolute value. Therefore, since $\frac{\partial z_{y_v}(\theta+\alpha\Delta)}{\partial\alpha} = \sum_{j\in J; v\in S(j)} \Delta_j f_j(y_v, x^n_{N_v})$

$$\begin{array}{rcl} |\frac{\partial\eta^{n,v}_{k,l}(\theta^*+\alpha\Delta, x^{(n)}_{N_v})}{\partial\alpha}| & \leq & \sum_{y_v\in I_v\setminus\{0\}} \frac{\partial z_{y_v}}{\partial\alpha} = \sum_{y_v\in I_v\setminus\{0\}} \sum_{j\in J; v\in S(j)} \Delta_j f_j(y_v, x^n_{N_v}) \\ & = & \sum_{j\in J; v\in S(j)} \Delta_j \sum_{y_v\in I_v\setminus\{0\}} f_j(y_v, x^n_{N_v}) = \langle\Delta, W^n\rangle \, . \end{array} \tag{7.17}$$

The Taylor series expansion of $\eta^{n,v}_{k,l}(\theta^* + \alpha\Delta, x^{(n)}_{N_v})$ yields

$$\eta^{n,v}_{k,l}(\theta^* + \alpha\Delta, x^{(n)}_{N_v}) = \eta^{n,v}_{k,l}(\theta^*, x^{(n)}_{N_v}) + \alpha\frac{\partial\eta^{n,v}_{k,l}(\theta^* + \alpha'\Delta, x^{(n)}_{N_v})}{\partial\alpha}, \;\; \alpha^{'} \in [0, \alpha] \, .$$

Let $K(\theta^* + \alpha^{'}\Delta, x^{(n)}_{N_v})$ denote the $d_v \times d_v$ matrix with entry $\frac{\partial\eta^{n,v}_{k,l}(\theta^*+\alpha\Delta, x^{(n)}_{N_v})}{\partial\alpha}$. Coming back to (7.16), we have

$$\begin{array}{rcl} k^{''}(\theta^* + \alpha\Delta) & = & \frac{1}{N}\sum_{n=1}^N \left[H(\theta^* + \alpha\Delta, x^{(n)}_{N_v}) \circ [W^n(W^n)^t]\right] \\ & = & \frac{1}{N}\sum_{n=1}^N H(\theta^*, x^{(n)}_{N_v}) \circ [W^n(W^n)^t] + \alpha\frac{1}{N}\sum_{n=1}^N K(\theta^* + \alpha^{'}\Delta, x^{(n)}_{N_v}) \circ [W^n(W^n)^t] \, . \end{array}$$

We write $||X||_2 = \lambda_{max}(X)$ for the operator norm of a matrix X. We recall that the Hadamard product of two positive semidefinite matrices is positive semidefinite and therefore

$$\lambda_{min}(K(\theta^* + \alpha^{'}\Delta, x_{N_v}^{(n)}) \circ [W^n(W^n)^t]) \geq -\lambda_{max}(K(\theta^* + \alpha^{'}\Delta, x_{N_v}^{(n)}) \circ [W^n(W^n)^t]).$$

Then since $|\alpha| < 1$, we have

$$\begin{aligned} q &= \min_{\alpha\in[0,1]} \lambda_{min}[\tfrac{1}{N}\textstyle\sum_{n=1}^{N} H(\theta^* + \alpha\Delta, x_{N_v}^{(n)})W^n(W^n)^t] \\ &\geq \lambda_{min}(\tfrac{1}{N}\textstyle\sum_{n=1}^{N}\left[H(\theta^*, x_{N_v}^{(n)}) \circ (W^n(W^n)^t)\right]) \\ &\qquad - \max_{\alpha\in[0,1]} ||\alpha\tfrac{1}{N}[\textstyle\sum_{n=1}^{N} K(\theta^* + \alpha\Delta, x_{N_v}^{(n)}) \circ (W^n(W^n)^t)]||_2 \\ &\geq C_{min} - \max_{\alpha\in[0,1]} \|\underbrace{\frac{1}{N}\sum_{n=1}^{N} \Delta^t W^n(W^n(W^n)^t)}_{A}\|_2 \\ &\geq C_{min} - \max_{\alpha\in[0,1]} ||A||_2 \ , \end{aligned} \tag{7.18}$$

where the last but one inequality is due to our Assumption (B). We now need to bound the spectral norm of $A = \frac{1}{N}\sum_{n=1}^{N} \Delta^t W^n(W^n(W^n)^t)$. For any $\alpha \in [0,1]$ and $y \in R^{d_v}$ with $||y||_F = 1$, we have

$$\begin{aligned} \langle y, Ay\rangle &= \frac{1}{N}\sum_{n=1}^{N}(\Delta^t W^n)(y^t W^n)^2 \leq \frac{1}{N}\sum_{n=1}^{N}|\Delta^t W^n|(y^t W^n)^2, \\ |\Delta^t W^n| &\leqslant \sqrt{d}||\Delta||_F = \sqrt{d}\delta \ . \end{aligned} \tag{7.19}$$

and, by definition of the operator norm and from Assumption (B),

$$\frac{1}{N}\sum_{n=1}^{N}(y^t W^n)^2 \leq ||\frac{1}{N}\sum_{n=1}^{N} W^n(W^n)^t||_2 < D_{max} \ . \tag{7.20}$$

From (7.18), (7.19) and (7.20), we obtain $\max_{\alpha\in[0,1]} ||A||_2 \leq D_{max}\sqrt{d}\delta$ and therefore

$$q \geq C_{min} - D_{max}\sqrt{d}\delta \ .$$

Substituting this into (7.16), we get

$$Q_2 \geq \frac{1}{2}\delta^2(C_{min} - D_{max}\sqrt{d}\delta). \tag{7.21}$$

From the two inequalities (7.15) and (7.21), it follows that

$$Q(\Delta) \geq Q_2 - |Q_1| \geq \frac{1}{2}\delta^2(C_{min} - D_{max}\sqrt{d}\delta) - \epsilon\sqrt{d}\delta. \tag{7.22}$$

To simplify the problem, we can choose δ such that $C_{min} - D_{max}\sqrt{d}\delta \geq \frac{C_{min}}{2}$, that is, $\delta \leq \frac{C_{min}}{2D_{max}\sqrt{d}}$. Then inequality (7.22) becomes

$$Q(\Delta) \geq \frac{C_{min}\delta^2}{4} - \epsilon\sqrt{d}\delta$$

and $Q(\Delta)$ is positive if we let $\delta = \frac{5\sqrt{d}\epsilon}{C_{min}}$. Moreover $\delta \leq \frac{C_{min}}{2D_{max}\sqrt{d}}$ yields the following bound of ϵ:

$$\epsilon \leq \frac{C_{min}^2}{10D_{max}d}.$$

We have therefore shown that (7.13) holds for $\delta = \frac{5\sqrt{d}\epsilon}{C_{min}}$ and the theorem is proved. □ In the next lemma, we will make use of Hoeffding inequality (see Hoeffding (1963), Theorem 2) which states the following. If If $X_1, X_2, \cdots, X_n$ are independent and $a_i \leq X_i \leq b_i (i = 1, 2, \cdots, n)$, then for $\epsilon > 0$

$$p(|\bar{X} - \mu| \geq \epsilon) \leq 2\exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 7.2 *Let $t_{J^{PS_v}}, k^{PS_v}$ and d_v be as defined above. For any $\epsilon > 0$, we have*

$$p(\{\|t_{J^{PS_v}} - (k^{PS_v})'(\theta^{v,*})\|_\infty \geq \epsilon\}) \leq 2d_v \exp(-\frac{N\epsilon^2}{2}) \tag{7.23}$$

and we also have

$$p(\{\max_{v\in V}\|y^v - (k^{PS_v})'(\theta^*)\|_\infty \geq \epsilon\}) \leq 2\sum_{v\in V} d_v \exp(-\frac{N\epsilon^2}{2}) \tag{7.24}$$

Proof. In this proof, as in the previous proof, we drop the superscripts v, PS except when necessary. For $j \in J^{PS_v}$, we clearly have

$$E_{\theta^*}\Big(\frac{\partial l(\theta)}{\partial \theta_j}\Big) = E_{\theta^*}\Big(t_j - \frac{\partial k(\theta)}{\partial \theta_j}\Big) = E_{\theta^*}\Big(\frac{1}{N}\sum_{n=1}^N f_j(x_v^{(n)}, x_{N_v}^{(n)}) - p(x_v = j_v|x_{N_v}^n) f_j(x_v = j_v, x_{N_v}^{(n)})\Big) = 0$$

and $|f_j(x_v^{(n)}, x_{N_v}^{(n)}) - p(x_v = j_v|x_{N_v}^{(n)})f_j(x_v = j_v, x_{N_v}^{(n)})| \leq 1$. Moreover, by Hoeffding inequality,

$$p(|t_j - k_j^{'}(\theta^*)| \geq \epsilon) \leq 2\exp -\frac{2N^2\epsilon^2}{2^2N} = 2\exp(-\frac{N\epsilon^2}{2})$$

Applying a union bound, we get, for each $v \in V$,

$$p(\|t_{J^{PS_v}} - k^{'}(\theta^*)\|_\infty \geq \epsilon) \leq \sum_{j=1}^{d_v} p(|t_j - k_j^{'}(\theta^*)| \geq \epsilon) \leq 2d_v \exp(-\frac{N\epsilon^2}{2})$$

from which it follows that

$$p\{\max_{v\in V}\|t_{J^{PS_v}} - (k^{PS_v})^{'}(\theta^*)\|_\infty \geq \epsilon\} \leq \sum_{v\in V} p(\|t_{J^{PS_v}} - (k^{PS_v})^{'}(\theta^*)\|_\infty \geq \epsilon) \leq 2\sum_{v\in V} d_v \exp(-\frac{N\epsilon^2}{2})$$

□

Proof of Theorem 5.1 Proof. Let $\epsilon = C\sqrt{\frac{\log p}{N}}$, where C is a constant that we will choose later in this proof. From Lemma 7.2, we have

$$p(\max_{v\in V}\|t_{J^{PS_v}} - (k^{PS_v})'(\theta^*)\|_\infty \geq C\sqrt{\frac{\log p}{N}}) \leq 2\sum_{v\in V} d_v \exp(-\frac{N(C\sqrt{\frac{\log p}{N}})^2}{2}) = \frac{2\sum_{v\in V} d_v}{p^{\frac{C^2}{2}}}$$

From Lemma 7.1, for $\epsilon = C\sqrt{\frac{\log p}{N}} \leq \frac{C_{min}^2}{10D_{max}d_v}$, i.e. for $N \geq (\frac{10CD_{max}d_v}{C_{min}^2})^2 \log p$, we have

$$\|t_{J^{PS_v}} - (k^{PS_v})'(\theta^*)\|_\infty \leq \epsilon \leq \frac{C_{min}^2}{10D_{max}d_v} \Rightarrow \|\hat{\theta}^{PS_v} - \theta^{v,*}\|_F \leq \frac{5\sqrt{d_v}\epsilon}{C_{min}}.$$

The global composite mle $\hat{\theta}$ obtained by the local averaging of the $\hat{\theta}^{PS_v}$ from each conditional model can then be bounded as follows:

$$\begin{aligned}
\|\hat{\theta} - \theta^*\|_F &\leq \big(\textstyle\sum_{v\in V}\|\hat{\theta}^{PS_v} - \theta^{v,*}\|_F^2\big)^{\frac{1}{2}} \\
&\leq (\textstyle\sum_{v\in V}(\frac{5\sqrt{d_v}C\sqrt{\frac{\log p}{N}}}{C_{min}})^2)^{\frac{1}{2}} \\
&= \frac{5C}{C_{min}}\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}}
\end{aligned}$$

Therefore under the condition $N \geq \max_{v\in V}\ (\frac{10CD_{max}d_v}{C_{min}^2})^2 \log p$, the following holds

$$\begin{aligned}
&\max_{v\in V}\|t_{J^{PS_v}} - (k^{PS_v})'(\theta^*)\|_\infty \geq C\sqrt{\tfrac{\log p}{N}}) \Rightarrow \|t_{J^{PS_v}} - (k^{PS_v})'(\theta^*)\|_\infty \geq C\sqrt{\tfrac{\log p}{N}}) \\
\Rightarrow\ &\|\hat{\theta}^{PS_v} - \theta^{v,*}\|_F \leq \tfrac{5\sqrt{d_v}\epsilon}{C_{min}} \Rightarrow \|\hat{\theta} - \theta^*\|_F \leq \tfrac{5C}{C_{min}}\sqrt{\tfrac{\sum_{v\in V} d_v \log p}{N}},
\end{aligned}$$

with

$$p(\|\hat{\theta}-\theta^*\|_F \leq \frac{5C}{C_{min}}\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}}) \geq p(\max_{v\in V}\|t_{J^{PS_v}}-k_v^{'}(\theta^*)\|_\infty \leq C\sqrt{\frac{\log p}{N}}) \geq 1-\frac{2\sum_{v\in V} d_v}{p^{\frac{C^2}{2}}}.$$

The theorem would make no sense if probability of the convergence rate was negative and thus C must satisfy

$$1-\frac{2\sum_{v\in V} d_v}{p^{\frac{C^2}{2}}} > 0 \Rightarrow C \geq \sqrt{2\frac{\log 2\sum_{v\in V} d_v}{\log p}}\ .$$

□